**From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Psychology

2012

# Behavioral, Neural, and Computational Investigations of Task Representations

Todd, Michael Travis

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01cz30ps68c

*U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf*

# Behavioral, Neural, and Computational Investigations of Task Representations

Michael Travis Todd

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Psychology
Adviser: Jonathan D. Cohen

June 2012

© Copyright by Michael Travis Todd, 2012.

All Rights Reserved

# Abstract

Context-sensitive stimulus-response mappings, which I call task representations, are a crucial part of flexible, goal-oriented behavior. Guided Activation Theory (GAT) and the related Gating models integrate many ideas about the behavioral and neural properties of task representations within computational frameworks. In this dissertation, I present several projects that test and extend these ideas. In the first project, I follow up on recent Gating model work, introducing novel connections with reinforcement learning (RL) theory that lead to a greatly simplified version of the previous model as well as to new insights about the limited capacity of task representations. In the second project, I investigate a previously proposed normative account of task representation updating (in the domain of task switching experiments) in terms of an effort-performance tradeoff. Converging theoretical and experimental work from this project reveals that a simple speed-accuracy tradeoff best explains the data, suggesting that the proposed effort-performance tradeoff is not a valid account of the findings. The final project attempts to measure context representations, a component of task representations, in human fMRI data. This project presents novel methodological developments (model-based multi-voxel pattern analaysis), and describes the application of the novel and existing methods to two new fMRI experiments. Surprisingly, neither experimental dataset yields measurable context representations using methods claimed by others to have produced success, although control analyses are successful.

# Acknowledgements

In graduate school, one ideally learns how to think, alone and especially in concert with others. This takes time: for me, it took five years, ten months, and one day. Yet, despite the inevitable bumps along the way, I never tired of this process, due entirely to my being surrounded by truly exceptional and supportive people.

From our first meeting onward, Jon Cohen has been so ceaseless a font of intellectual vigor, so generous with resources and praise, so nuanced in modulating frankness with diplomacy, and so unfailingly optimistic about my education, that it is difficult to thank him adequately without becoming maudlin. My most succinct description of Jon as a scientist is that he perfectly resolves the tension between theory and experiment that lies at the heart of cognitive neuroscience. I would imitate him if I could, and studying under him has been a great privilege.

As repository of much of the lab's methodological knowledge, and as oracle for the often unavailable and occasionally inscrutable Jon, Leigh Nystrom is the go-to for confused or frustrated students and postdocs, a burden that he seems to shoulder effortlessly. I am thankful for the many hours he spent listening about my problems and advising me on neuroimaging analyses.

After my time at Princeton, I will approach any technical problem with confidence, and I owe much to Yael Niv for setting me on this path. Yael taught me how to take math out of the classroom and actually *use* it as the natural approach to real problems. In a field where "simple" models are sequences of nonlinear functions, teaching this lesson required substantial conviction, but Yael was equal to the task. This began with making me interact with technical literature at the level of completing proofs, and I clearly remember working through the first one — computing the gradient for Williams's (1992) REINFORCE algorithm — with Yael sitting beside me, my hand actually shaking with uncertainty. She patiently coaxed me through that exercise and pointed me to relevant follow-ups, and from there I have never looked back.

Her energetic and uncompromising rigor has profoundly shaped and sharpened my thinking.

The intellectual environment of this department has been, for me, essentially collaborative, and I learned much about how to think *with* people during many excellent group meetings led by a recurring cast of professors. Jon Cohen, Yael Niv, Ken Norman, Matt Botvinick, Carlos Brody, Phil Holmes, and Naomi Leonard have all been models of how to approach problems collaboratively with multiple scholars from multiple disciplines. I have many fond memories from these meetings, and I am anxious to find a new home where such lively collaborations occur.

My education came in no small part from discussions with postdocs and fellow students. Chand Chandrasekaran, Sunny Khemlani, Joe McGuire, Bob Wilson, Damon Tomlin, Mina Cikara, Chris Said, Anna Schapiro, and José Fernandes all provided me with many hours of edifying conversation on every aspect of our field. Of course the most frequent, indeed incessant peer interactions were with officemates, and I am accordingly indebted to Sam Gershman and John White for sharing their occasionally scintillating, always substantive thoughts of everything from New York's avant-garde scene to what it means to search for truth in data.

Even for graduate students, there is much to be said about six years of one's life that has little to do with education. When I was not sitting in front of a computer (and occasionally when I was), I was usually fortunate enough to be in the company of friends and family. Chand and Sunny taught me how to cook dosas multiple times, lessons that I appreciated despite promptly forgetting the skill each time. As Princeton was far from my family, I was lucky that Joe and Mary took me in like their own bachelor brother, and their refrigerator was never closed against me. Anuj and Julia were similarly wonderful hosts of fun evenings (and excellent cooks). Katie K. provided many evenings of Rock Band and cupcakes. Although there was no gastronomy at my house, I remember fondly the many nights that Mina and I stayed

up watching trashy television and talking shop until 4am, as well as many nights when we stayed up just as late with visitors, barbecuing and playing boardgames. I was blessed to find such wonderful friends here at Princeton. Last but certainly not least, one of the best parts of grad school is leaving to go home on occasion. It was wonderful to see my siblings begin new families of their own during my time here, and it was essential to my sanity that I was welcomed back home for luxuriously indulgent and often lengthy visits. Leaving Princeton would be bittersweet if I were not returning home to live near my loving and supportive family.

"Theory, friend, is old and grey,

And green the golden tree of life!"

— Mephistopheles in *Faust I* (Goethe, 1808)

"We're not going to figure this stuff out with *word* models!"

— Jon Cohen

# Contents

Abstract . . . . . iii

Acknowledgements . . . . . iv

**1 Introduction** **1**

1.1 Revising the Gating model . . . . . 3

1.2 Rational Analysis of Task Switching . . . . . 5

1.3 Mapping context representations with fMRI: Method and experimental study . . . . . 7

1.4 Conclusions . . . . . 8

**2 Learning Dynamics: The Gating Model** **9**

2.1 Introduction . . . . . 9

2.2 Model Architecture . . . . . 12

2.3 Benchmark Performance and Psychological Data . . . . . 14

2.3.1 12-AX Performance . . . . . 15

2.3.2 Behavioral Data . . . . . 16

2.4 Discussion . . . . . 19

2.4.1 Relation to Other Theoretical Work . . . . . 19

2.4.2 Implications for Working Memory and Cognitive Control . . . 21

2.4.3 Psychological & Neural Validity . . . . . 22

**3 Within-Trial Dynamics: Model and Behavioral Experiment** **23**

3.1 Introduction . . . 23
3.1.1 Task Switching . . . 23
3.1.2 Motivational Effects in Task Switching . . . 26
3.2 The Model . . . 27
3.2.1 Task Switching Experiment Structure . . . 27
3.2.2 Rational Analysis . . . 30
3.2.3 Exploring Model Behavior in Pre-Existing Experiments . . . 34
3.2.4 Model Summary . . . 41
3.3 Testing the Models: Deadlined Task Switching Paradigm . . . 42
3.3.1 Experiment 1 . . . 42
3.3.2 Experiment 2 . . . 46
3.4 General Discussion . . . 49
3.4.1 Motivational effects in task switching . . . 51
3.4.2 Model Assumptions & Alternatives . . . 52
3.4.3 Conclusion . . . 53
3.5 Supporting Mathematical Details . . . 53
3.5.1 Rational Analysis of Basic Task Switching . . . 53
3.5.2 Special Cases . . . 63
3.5.3 Alternate Optimality Criterion: Undiscounted Expected Future Reward . . . 67
3.5.4 Obtaining My DDM Equations From Ratcliff (1978) . . . 69

**4 Confounds in fMRI Mapping of Context Representations 71**

4.1 Introduction . . . 71
4.2 Directional vs Directionless Single-Subject Summaries . . . 75
4.3 Experimental Demonstration: Rule Representations . . . 82
4.3.1 Methods . . . 83
4.3.2 Results . . . 88

4.3.3 fMRI Results . . . . . 90
4.3.4 Experiment Discussion . . . . . 93
4.4 General Discussion . . . . . 97

**5 Increasing Sensitivity and Interpretability of fMRI Mapping with Model-Based Multivoxel Pattern Analysis 100**

5.1 Introduction . . . . . 100
5.1.1 Model-based MVPA . . . . . 101
5.1.2 Generative classifiers and model-based MVPA . . . . . 102
5.1.3 Fluctuations in Task Representations . . . . . 104
5.1.4 GAT as a Generative Model of Context Fluctuations . . . . . 105
5.2 Modeling Introduction: The GNB Classifier . . . . . 109
5.2.1 Generative Model Specification . . . . . 109
5.2.2 Parameter Estimation . . . . . 111
5.2.3 Model Inversion and Cross-Validation . . . . . 114
5.3 Model 0: Basic Mixture . . . . . 116
5.3.1 Generative Model Specification . . . . . 117
5.3.2 Parameter Estimation . . . . . 118
5.3.3 Model Inversion and Cross-validation . . . . . 126
5.3.4 Simulations: Model 0 vs GNB . . . . . 128
5.3.5 Model 0 Discussion . . . . . 129
5.4 Model 1: Basic GAT . . . . . 130
5.4.1 Generative Model Specification . . . . . 130
5.4.2 Parameter Estimation . . . . . 133
5.4.3 Model Inversion and Cross-validation . . . . . 136
5.4.4 Model 1 Simulations . . . . . 138
5.4.5 Model 1 Discussion . . . . . 139
5.5 Model 2: Behavioral differences across cognitive states . . . . . 141

5.5.1 Generative Model Specification . . . . . 141
5.5.2 Parameter Estimation . . . . . 143
5.5.3 Model Inversion and Cross-validation . . . . . 145
5.5.4 Model 2 Discussion . . . . . 146
5.6 Additional Simulations . . . . . 146
5.7 Application to fMRI Data . . . . . 150
5.7.1 Methods . . . . . 150
5.7.2 Results . . . . . 150
5.8 General Discussion . . . . . 152
5.8.1 Cross-validation versus similarity analysis . . . . . 152
5.8.2 Connectionist Implementations of GAT . . . . . 154
5.8.3 Application to fMRI Data . . . . . 158
5.8.4 Conclusions . . . . . 159

**6 Discussion** **160**

# Chapter 1

# Introduction

One hallmark of realistic environments is context, or information that changes frequently and is not immediately observable at the time when decisions must be made, but is nonetheless crucial for pursuing goals. In psychological terms, the required stimulus-response mapping changes frequently in such environments, and thus navigating them necessitates flexible yet robust representation of the currently appropriate mapping in working memory. This component of working memory is known as a rule or task representation, and its study is central to understanding complex, goal-oriented behavior. Task representations have been consistently associated with prefrontal cortex (PFC Badre, Kayser, & D'Esposito, 2010; D'Esposito et al., 1995; Badre & Wagner, 2004; Bunge, 2004; Duncan, 2001; Petrides, 2005; Wallis, Anderson, & Miller, 2001; Montojo & Courtney, 2008; Cohen et al., 1997). Yet the wealth of empirical data regarding task representations is complex and resists obvious simplifications, highlighting the need for computational models in furthering our understanding.

Guided Activation Theory (GAT Miller & Cohen, 2001) is a family of models describing a particular relationship between PFC, posterior cortex and subcortical structures, and task representations. Cohen, Dunbar, and McClelland's (1990) model

of data from Stroop experiments (Stroop, 1935) was an early, purely cognitive precursor of GAT that specified the theory's characteristic cognitive features. This model posited that task representations emerge from interactions in a parallel distributed processing system in which context units allocate attention to (i.e., select) goal-relevant stimulus-response processing pathways from among parallel competing pathways. Thus, in complex environments in which many conflicting stimulus-response mappings may be simultaneously activated by bottom-up stimulus information, task representations emerge from the strategic selection of one mapping among many. Cohen and Servan-Schreiber (1992) subsequently linked the model's context representations to PFC in the domain of behavioral deficits associated with schizophrenia. Desimone and Duncan (1995) later proposed a model of visual attention with strong links between neural substrates and parallel competing representations, arguing that visual attention is an emergent property of competition among low-level representations in visual cortex, and that PFC was somehow involved in biasing this competition toward goal-relevant information. Following up on these earlier works, Miller and Cohen (2001) definitively articulated GAT, positing that the function of PFC is to actively maintain context representations and to use this to bias competition in posterior cortical and subcortical structures toward goal-relevant information. Thus GAT frames task representations as emerging from a sort of internal attention (analogous to visual attention) to goal-relevant information-processing pathways, guided by actively maintained contextual information in the PFC.

One challenge for GAT is the stability-flexibility paradox. That is, the model requires task representations to be both stable in the face of goal-irrelevant information, and yet also flexible enough to be changed or updated by new, goal-relevant information. The Gating model is an extension of GAT that addresses this paradox by positing a relationship between task representations and dopaminergic systems presumed to be involved in reinforcement learning (RL). Initial formulations of the

Gating model assumed that phasic bursts from midbrain dopaminergic nuclei (Braver & Cohen, 1999, 2000) directly "gated" goal-relevant information into PFC where it would be actively maintained until the next dopaminergic burst signal. Later formulations posited a more complex relationship, with phasic dopamine bursts training a gating "actor" in the basal ganglia that then gated goal-relevant information into PFC via corticostriatal loops (Frank, Loughry, & O'Reilly, 2001; O'Reilly & Frank, 2006). Although there is not yet sufficient data to adjudicate between the details of the earlier and later forms of the Gating model, there is a key commonality between them. Specifically, both forms solve GAT's stability-flexibility paradox by assuming that basic RL learning processes identify goal-relevant information and cause it to be gated into PFC. This escapes the trap of introducing another homunculus that must control the RL system in turn, because the *sine qua non* of this system is its ability to learn complex behaviors (e.g., goal-oriented gating) from primordial reward and punishment stimuli. This effectively extends GAT's distributed task representations to include: context representations in PFC, stimulus-response representations in posterior cortex and subcortical structures, and context updating policies in the dopaminergic RL system.

Figure 1.1 summarizes GAT and the Gating models. In this dissertation, I present theoretical as well as behavioral and neuroimaging studies that further develop and test aspects of these ideas. I introduce each of these studies in the following sections.

## 1.1 Revising the Gating model

This project followed up on the Gating model proposed by O'Reilly and Frank (2006)[1], emphasizing a previously unexplored link between the Gating model and a particular

[1]The title of that work begins "Making Working Memory Work...," and I continue the more general emphasis on working memory in this chapter. It is left implicit that context and task representations are a specific type of working memory, and this project therefore addresses these more specific representations as well.

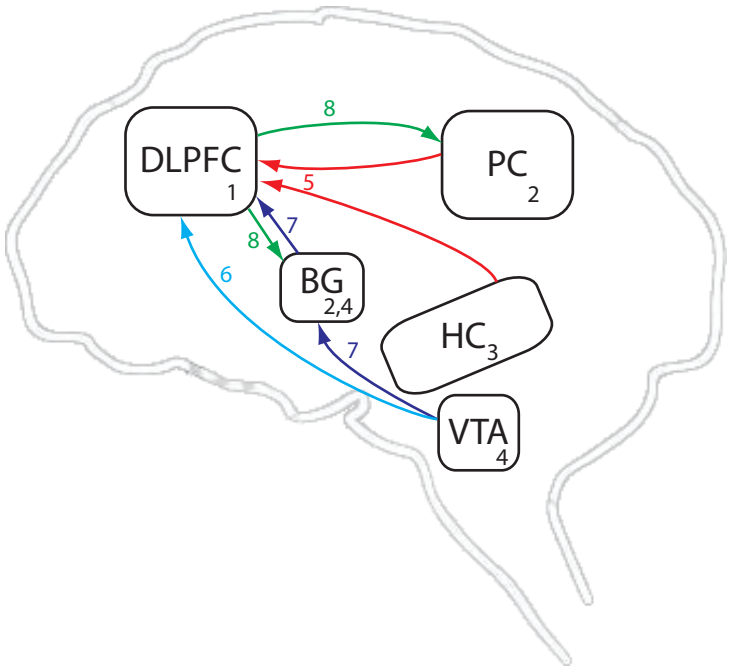



Figure 1.1: Assumptions of GAT and the Gating models, including relevant aspects of neural RL models. Task representation cognitive structure (without neural links) first proposed by Cohen, Dunbar, and McClelland (1990). 1: Context representations are actively maintained in DLPFC (Cohen & Servan-Schreiber, 1992; Miller & Cohen, 2001). 2: Familiar stimulus-response mappings are represented in posterior cortex (PC: Miller & Cohen, 2001) and basal ganglia (BG: Joel, Niv, & Ruppin, 2002; Frank et al., 2001; O'Reilly & Frank, 2006). 3: Novel (e.g., "indirect": Cohen, Dunbar, & McClelland, 1990) stimulus-response mappings may be represented in hippocampus (HC: Miller & Cohen, 2001). 4: RL signals are carried by phasic bursts from dopamine (DA) neurons in ventral tegmental area (VTA) and substantia nigra (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). These train policies and values into BG (Houk, Adams, & Barto, 1995; Reynolds, Hyland, & Wickens, 2001; Reynolds & Wickens, 2002). 5: Context cues from the environment as well as HC project into DLPFC (Braver & Cohen, 1999; Frank et al., 2001; O'Reilly & Frank, 2006). 6: RL signal mediated by phasic DA directly gates only task-relevant context cues into DLPFC (Braver & Cohen, 1999, 2000). 7: RL trains gating policy into BG, which then gates only task-relevant context cues into DLPFC (Frank et al., 2001; O'Reilly & Frank, 2006). 8: DLPFC selects among competing mappings in PC and BG via top-down attentional bias (Cohen & Servan-Schreiber, 1992; Desimone & Duncan, 1995; Miller & Cohen, 2001).

branch of RL theory. That is, the elements of RL theory that have typically been used to derive neuroscience models (e.g., Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997) concern decision making in environments that, in contrast to the "realistic" environments posed above, are by definition context-free, requiring no actively maintained task representations. This is the theory of decision making in so-called Markov Decision Processes (Puterman, 1994; Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998). However, a different area of RL theory, less widely known among neuroscientists, concerns decision making in more complex, realistic environments that do require working memory. This is the theory of decision making in Partially Observable Markov Decision Processes (e.g., Kaelbling, Littman, & Cassandra, 1998). In this project, I established a link between this branch of RL theory and the Gating model.

The benefit of establishing this link was twofold. First, this led to a highly formal definition of working memory in general in terms of the internal memory device that theoretical decision making agents must use in order to make effective decisions in partially observable environments. Second, it led to the recognition that the biologically inspired but highly complex form of the Gating model proposed by O'Reilly and Frank (2006) could be simplified to a three-parameter model while preserving the key concepts (Todd, Niv, & Cohen, 2008). The new, simpler Gating model was useful for exploring how the definition of working memory in terms of reward maximization and RL could be used to explain the characteristic limitations of the working memory system (Cleeremans & McClelland, 1991; Cowan, 2000).

## 1.2 Rational Analysis of Task Switching

This project used the task switching literature to examine task representation dynamics at the within-trial timescale. A central finding in that literature is the residual

switch cost, or behavioral deficit from switching from one task to another that remains even after ample preparation time (e.g., Allport, Styles, & Hsieh, 1994; Meiran, 1996; Rogers & Monsell, 1995). While the residual switch cost was originally interpreted in terms of cognitive structural factors, later work suggested that the residual switch cost is in fact "optional" for participants, reflecting motivational factors, specifically an effort-performance tradeoff (De Jong, 2000; Nieuwenhuis & Monsell, 2002). In this project I used rational analysis theory (Anderson, 1990, 1991) to articulate these two alternative interpretations as competing models within a formal framework. Critically, this afforded examination of variance attributable to a motivation-sensitive effort-performance tradeoff after precisely controlling for variance associated with a motivation-sensitive speed-accuracy tradeoff.

The use of rational analysis here is akin to implementing a Gating model of task switching at a high level of abstraction, in that both approaches assume that task representations are updated in such a way as to maximize rewards (and minimize costs). The rational analysis approach was chosen as it more readily exposed the effort-performance tradeoff without explicit learning assumptions that are necessary for RL accounts (including Gating models) but that do not appear to correspond well with task switching experiments.

Simulations showed the validity of the overall framework (i.e., both competing models) with regard to existing task switching data: the optimal resolution of the effort-performance and/or speed-accuracy tradeoffs leads to behavior that is consistent with several reported phenomena, including residual switch costs (Rogers & Monsell, 1995), the failure to engage pattern (De Jong, 2000; Nieuwenhuis et al., 2002), and probabilistic cueing effects (Dreisbach, Haider, & Kluwe, 2002; Dreisbach & Haider, 2006). To adjudicate between the two models, I developed a behavioral experiment paradigm (deadlined task switching) in which performance costs could be parametrically manipulated. Data from two experiments using variants of this

paradigm produced results similar to those previously reported by De Jong (2000) and Nieuwenhuis and Monsell (2002). Nonetheless, the data from both experiments failed to support the effort-performance tradeoff after controlling for the speed-accuracy tradeoff. This suggests a reinterpretation of the data that originally led De Jong (2000) and Nieuwenhuis et al. (2002) to suggest the effort-performance tradeoff explanation of the residual switch cost. Those data now appear to be best explained by a speed-accuracy tradeoff alone, and it does not appear that the within-trial task representation dynamics indexed by the residual switch cost are interpretable in terms of motivational factors in the manner previously proposed.

## 1.3 Mapping context representations with fMRI: Method and experimental study

The goal of this project was to map the neural substrates of context representations, the part of task representations thought to reside within PFC, using fMRI. Mapping these neural substrates is an important avenue for testing various aspects of the models described above. This project comprises two chapters in which I describe my efforts to map context representations.

In the first chapter, I describe a confound that is associated with a type of fMRI analysis, but that is not widely recognized. Given this confound and newly collected fMRI data, I argue that recently reported fMRI maps of context representations (Bode & Haynes, 2009; Woolgar, Thompson, Bor, & Duncan, 2011; Reverberi, Görgen, & Haynes, 2011; Cole, Etzel, Zacks, Schneider, & Braver, 2011) most likely reflect maps of mental effort or difficulty, a variable that is confounded with context representations at the within-subject level. I thus conclude that existing fMRI mapping studies of context representations have yielded relatively little success compared to electrophysiological investigations of context representations (e.g., Rainer, Rao, & Miller, 1999;

Romo, Brody, Hernández, & Lemus, 1999) or fMRI mapping of other domains such as visual representations (e.g., Haxby et al., 2001; Kamitani & Tong, 2005; Haynes & Rees, 2005).

One possible reason is that standard fMRI analysis methods may be particularly ill-suited for detecting (and thus mapping) context representations. To address this possibility, in the second chapter I developed a new fMRI analysis method that incorporates aspects of GAT directly into an existing multivoxel pattern-analysis (MVPA) tool. The resulting model-based MVPA tool is theoretically more sensitive to detecting context representations and more interpretable (i.e., less confounded) than standard methods. This is demonstrated via simulations. I then re-analyze our fMRI data with the newly developed model-based MVPA methods as well as existing methods. Unfortunately, the model-based MVPA variants that I developed did not reveal robust context representation maps. Nonetheless, I argue that model-based MVPA is of substantial theoretical interest, and should be pursued further.

## 1.4 Conclusions

The projects described above comprise a related body of theoretical and experimental studies on the behavioral properties and neural substrates of task representations. The theoretical work advances previous modeling work, yielding new predictions about the relationship between task representations, behavior, and neural measurements. The experiments correspondingly employed model-driven designs and methods to test these predictions.

# Chapter 2

# Learning Dynamics: The Gating Model

A slightly different version of this chapter was in published as Todd, M., Niv, Y., & Cohen, J.D. (2008). Learning to use working memory in partially observable environments through dopaminergic reinforcement. In D. Koller, Y. Bengio, D. Schuurmans, L. Bottou & A. Culotta (Eds.), Advances in neural information processing systems (Vol. 21).

## 2.1 Introduction

Working memory is loosely defined in cognitive neuroscience as information that is (1) internally maintained on a temporary or short term basis, and (2) required for tasks in which immediate observations cannot be mapped to correct actions. It is widely assumed that prefrontal cortex (PFC) plays a role in maintaining and updating working memory. However, relatively little is known about how PFC develops useful working memory representations for a new task. Furthermore, current work focuses on describing the structure and limitations of working memory, but does not ask why, or in what general class of tasks, is it necessary. Borrowing from the theory

of optimal control in partially observable Markov decision problems (POMDPs), we frame the psychological concept of working memory as an internal state representation, developed and employed to maximize future reward in partially observable environments. We combine computational insights from POMDPs and neurobiologically plausible models from cognitive neuroscience to suggest a simple reinforcement learning (RL) model of working memory function that can be implemented through dopaminergic training of the basal ganglia and PFC.

The Gating framework (Braver & Cohen, 1999, 2000; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005; O'Reilly & Frank, 2006) is a series of cognitive neuroscience models developed to explain how dopaminergic RL signals can shape useful working memory representations. Computationally this framework models working memory as a collection of past observations, each of which can occasionally be replaced with the current observation, and addresses the problem of learning when to update each memory element versus maintaining it. In the original Gating model (Braver & Cohen, 1999, 2000) the PFC contained a unitary working memory representation that was updated whenever a phasic dopamine (DA) burst occurred (e.g., due to unexpected reward or novelty). That model was the first to connect working memory and RL via the temporal difference (TD) model of DA firing (Montague et al., 1996; Schultz et al., 1997), and thus to suggest how working memory might serve a normative purpose. However, that model had limited computational flexibility due to the unitary nature of the working memory (i.e., a single-observation memory controlled by a scalar DA signal). More recent work (Rougier et al., 2005; O'Reilly & Frank, 2006) has partially repositioned the Gating framework within the Actor/Critic model of mesostriatal RL (Houk et al., 1995; Joel, Niv, & Ruppin, 2002), positing memory updating as but another cortical action controlled by the dorsal striatal "actor." This architecture increased computational flexibility by introducing multiple working memory elements, corresponding to multiple corticostriatal loops, that could be quasi-independently

updated. However, that model combined a number of components (including supervised and unsupervised learning, and complex neural network dynamics), making it difficult to understand the relationship between simple RL mechanisms and working memory function. Moreover, because the model used the Rescorla-Wagner-like PVLV algorithm (O'Reilly & Frank, 2006) rather than TD (Montague et al., 1996; Schultz et al., 1997) as the model of phasic DA bursts, the model's behavior and working memory representations were not directly shaped by standard normative criteria for RL models (i.e., discounted future reward or reward per unit time).

We present a new Gating model, synthesizing the mesostriatal Actor/Critic architecture of [4] with a normative POMDP framework, and reducing the Gating model to a four-parameter, pure RL model in the process. This produces a model very similar to previous machine learning work on "model-free" approximate POMDP solvers (McCallum, 1995; Peshkin, Meuleau, & Kaelbling, 1999), which attempt to form good solutions without explicit knowledge of the environment's structure or dynamics. That is, we model working memory as a discrete memory system (a collection of recent observations) rather than a continuous "belief state" (an inferred probability distribution over hidden states). In some environments this may permit only an approximate solution. However, the strength of such a system is that it requires very little prior knowledge, and is thus potentially useful for animals, who must learn effective behavior and memory-management policies in completely novel environments (i.e., in the absence of a world model). Therefore, we retain the computational flexibility of the more recent Gating models (Rougier et al., 2005; O'Reilly & Frank, 2006), while re-establishing the goal of defining working memory in normative terms (Braver & Cohen, 1999, 2000).

To illustrate the strengths and limitations of the model, we apply it to two representative working-memory tasks. The first is the 12-AX task proposed as a Gating benchmark in (O'Reilly & Frank, 2006). Contrary to previous claims that TD learning

is not sufficient to solve this task, we show that with an eligibility trace (i.e., TD($\lambda$) with $0 \leq \lambda \leq 1$), the model can achieve optimal behavior. The second task highlights important limitations of the model. Since our model is a POMDP solver and POMDPs are, in general, intractable (i.e., solution algorithms require an infeasible number of computations), it is clear that our model must ultimately fail to achieve optimal performance as environments increase even to moderate complexity. However, human working memory also exhibits sharp limitations. We apply our model to an implicit artificial grammar learning task (Cleeremans & McClelland, 1991) and show that it indeed fails in ways reminiscent of human performance. Moreover, simulating this task with increased working memory capacity reveals diminishing returns as capacity increases beyond a small number, suggesting that the "magic number" limited working memory capacity found in humans (Cowan, 2000) might in fact be optimal from a learning standpoint.

## 2.2 Model Architecture

As with working memory tasks, a POMDP does not admit an optimal behavior policy based only on the current observation. Instead, the optimal policy generally depends on some combination of memory as well as the current observation. Although the type of memory required varies across POMDPs, in certain cases a finite memory system is a sufficient basis for an optimal policy. Peshkin et al. (1999) used an external finite memory device (e.g., a shopping list) to improve the performance of RL in a model-free POMDP setting. Their model's "state" variable consisted of the current observation augmented by the memory device. An augmented action space, consisting of both memory actions and motor actions, allowed the model to learn effective memory-management and motor policies simultaneously. We integrate this approach with the Gating model, altering the semantics so that the external memory

device becomes internal working memory (presumed to be supported in PFC), and altering the Gating model so that the role of working memory is explicitly to support optimal behavior (in terms of discounted future reward) in a POMDP.

As with Peshkin et al. (1999), the key difference between our model and standard RL methods is that our state variable includes controlled memory elements (i.e., working memory), which augment the current observation. The action space is similarly augmented to include memory or gating actions, and the model learns by trial-and-error how to update its working memory (to resolve hidden states when such resolution leads to greater rewards) as well as its motor policy. The task for our model then, is to learn a working memory policy such that the current internal state (i.e., memory and current observation) admits an optimal behavioral policy.

Our model (Table 1) consists of a critic, a motor actor, and several gating actors. As in the standard Actor/Critic architecture, the critic learns to evaluate (internal) states and, based on the ongoing temporal difference of these values, generates at each time step a prediction error (PE) signal, thought to correspond to phasic bursts and dips in DA (Schultz et al., 1997). The PE is used to train the critic's state values and the policies of the actors. The motor actor also fulfills the usual role, choosing actions to send to the environment based on its policy and the current internal state. Finally, gating actors correspond one-to-one with each memory element. At each time point, each gating actor independently chooses (via a policy based on the internal state) whether to (1) maintain its element's memory for another time step, or (2) replace (update) its element's memory with the current observation.

To remain aligned with the Actor/Critic online learning framework of mesostriatal RL (Houk et al., 1995; Joel et al., 2002), learning in our model is based on REINFORCE (Williams, 1992) modified for expected discounted future reward as in Dayan and Abbott (2001), rather than the original Monte-Carlo policy learning algorithm in Peshkin et al. (1999) (which is more suitable for offline, episodic learn-

ing). Furthermore, because it has been shown that eligibility traces are particularly useful when applying TD to POMDPs (e.g., Singh, Jaakkola, & Jordan, 1994; Loch & Singh, 1998), we used TD($\lambda$), taking the characteristic eligibilities of the REINFORCE algorithm as the impulse function for a replacing eligibility trace (Sutton & Barto, 1998). For simplicity of exposition and interpretation, we used tabular policy and state-value representations throughout.

## 2.3 Benchmark Performance and Psychological Data

We now describe the model's performance on the 12-AX task proposed by O'Reilly and Frank (2006) as a benchmark for Gating models. We then turn to a comparison of the model's behavior against actual psychological data.

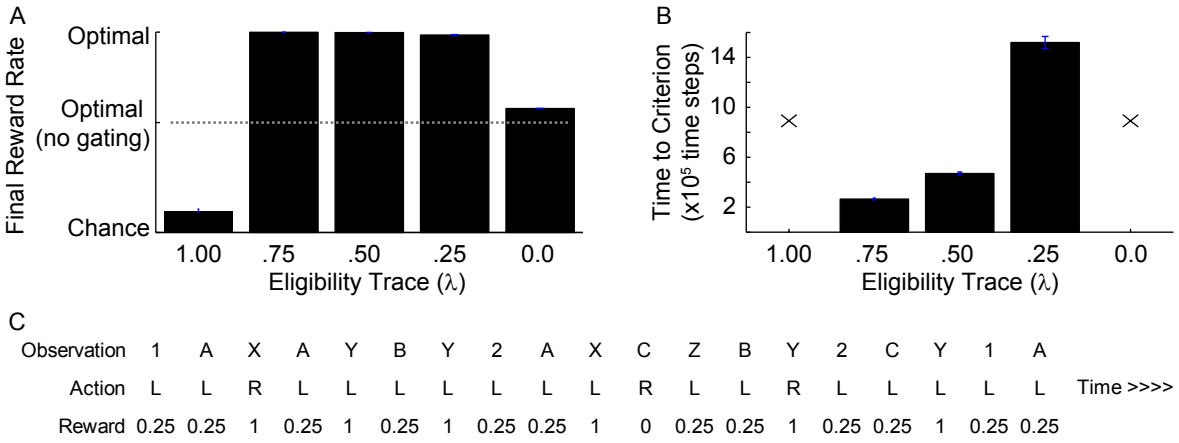



Figure 2.1: 12-AX: Average performance over 40 training runs, each consisting of 2107 timesteps. (A) As indicated by reward rate over the last 105 time steps, the model learns an optimal policy when the eligibility trace parameter, , is between zero and one. (B) The time required for the model to reach 300 consecutive correct trials increases rapidly as  decreases. (C) Sample sequence of the 12-AX task.

### 2.3.1 12-AX Performance

O'Reilly and Frank (2006) used the 12-AX task to illustrate the problem of learning a task in which correct behavior depends on multiple previous observations. In the task (Figure 1C), subjects are presented with a sequence of observations drawn from the set 1, 2, A, B, C, X, Y, Z. They gain rewards by responding L or R according to the following rules: Respond R if (1) the current observation is an X, the last observation from the set A, B, C was an A, and the last observation from the set 1, 2 was a 1; or (2) the current observation is a Y, the last observation from the set A, B, C was a B, and the last observation from the set 1, 2 was a 2. Respond L otherwise. In our implementation, reward is 1 for correct responses when the current observation is X or Y, 0.25 for all other correct responses, and 0 for incorrect responses.

We modeled this task using two memory elements, the minimum theoretically necessary for optimal performance. The results (Figure 1A,B) show that our TD($\lambda$) Gating model can indeed achieve optimal 12-AX performance. The results also demonstrate the reliance of the model on the eligibility trace parameter, $\lambda$, with best performance at high intermediate values of $\lambda$. When $\lambda$=0, the model finds a suboptimal policy that is only slightly better than the optimal policy for a model without working memory. With $\lambda$=1 performance is even worse, as can be expected for an online policy improvement method with non-decaying traces — a point of comparison with Peshkin et al. (1999) to which we will return in the Discussion. These results are consistent with previous work showing that TD(0) performs poorly in partially observable (non-Markovian) settings (Singh et al., 1994), whereas TD($\lambda$) (without memory) with $\lambda \approx 0.9$ performs best (Loch & Singh, 1998). Indeed, early in training, as our model learns to convert a POMDP to an MDP via its working memory, the internal state dynamics are not Markovian, and thus an eligibility trace is necessary.

### 2.3.2 Behavioral Data

We are the first to interpret the Gating framework (and the use of working memory) as an attempt to solve POMDPs. This brings a large body of theoretical work to bear on the properties of Gating models. Importantly, it implies that, as task complexity increases, both the Gating model and humans must fail to find optimal solutions in reasonable time frames due to the generally intractable nature of POMDPs. Given this inescapable conclusion, it is interesting to compare model failures to corresponding human failures: a pattern of failures matching human data would provide support for our model. In this subsection we describe a simulation of artificial grammar learning (Cleeremans & McClelland, 1991), and then offer an account of the pervasive "magic number" observations concerning limits of working memory capacity (e.g., Cowan, 2000).

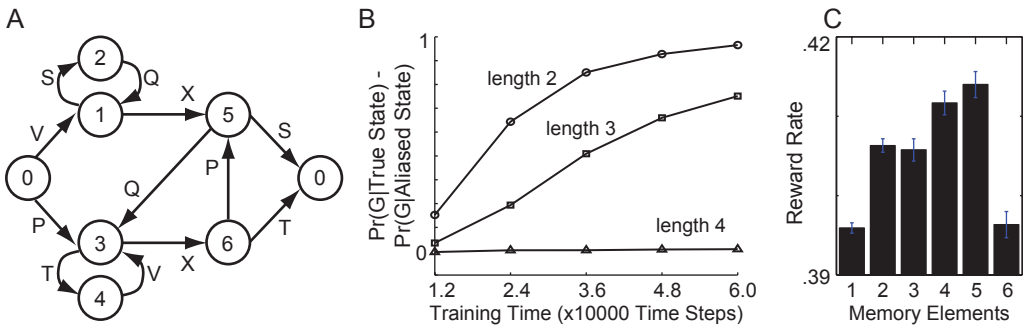



Figure 2.2: (A) Artificial grammar from [11]. Starting from node 0, the grammar generates a continuing sequence of observations. All nodes with two transitions (edges) make either transition with p=0.5. Edge labels mark grammatical observations. At each transition, the grammatical observation is replaced with a random, ungrammatical, observation with p=0.15. The task is to predict the next observation at each time point. (B) The model shows a gradual increase in sensitivity to sequences of length 2 and 3, but not length 4, replicating the human data. Sensitivity is measured as probability of choosing grammatical action for the true state, minus probability of choosing grammatical action for the aliased state; 0 indicates complete aliasing, 1 complete resolution. (C) Model performance (reward rate) averaged over training runs with variable numbers of time steps shows diminishing returns as the number of memory elements increases.

In artificial grammar learning, subjects see a seemingly random sequence of observations, and are instructed to mimic each observation as quickly as possible (or to predict the next observation) with a corresponding action. Unknown to the subjects, the observation sequence is generated by a stochastic process called a "grammar" (Figure 2A). Artificial grammar tasks constitute POMDPs: the (recent) observation history can predict the next observation better than the current observation alone, so optimal performance requires subjects to remember information distilled from the history. Although subjects typically report no knowledge of the underlying structure, after training their reaction times (RTs) reveal implicit structural knowledge. Specifically, RTs become significantly faster for "grammatical" as compared to "ungrammatical" observations (see Figure 2).

Cleeremans and McClelland (1991) examined the limits of subjects' capacity to detect grammar structure. The grammar they used is shown in Figure 2A. They found that, although subjects grew increasingly sensitive to sequences of length two and three throughout training, (as measured by transient RT increases following ungrammatical observations), they remained insensitive, even after 60,000 time steps of training, to sequences of length four. This presumably reflected a failure of subjects' implicit working memory learning mechanisms, and was confirmed in a second experiment. Our model replicated these results, as shown in Figure 2B. To simulate the task, we gave the model two memory elements (results were no different with three elements), and reward 1 for each correct prediction. We tested the model's ability to resolve states based on previous observations by contrasting its behavior across pairs of observation sequences that differed only in the first observation. State resolution based on sequences of length two, three, and four were represented by VS versus XS (leading to predictions Q vs. V/P, respectively), SQX versus XQX (S/Q vs. P/T), and XTVX versus PTVX (S/Q vs. P/T), respectively.

In this task, optimal use of information from sequences of length four or more

proved impossible for the model and, apparently, for humans. To understand intuitively this limitation, consider a problem of two hidden states, 1 and 2, with optimal actions L and R, respectively. The states are preceded by identical observation sequences of length n. However, at n+1 time steps in the past, observation A precedes state 1, whereas observation B precedes state 2. The probability that A/B are held in memory for the required n+1 time steps decreases geometrically with n, thus the probability of resolving states 1 and 2 decreases geometrically. Because the agent cannot resolve state 1 from state 2, it can never learn the appropriate 1-L, 2-R action preferences even if it explores those actions, a more insidious problem than an RL agent faces in a fully observable setting. As a result, the model cant reinforce optimal gating policies, eventually learning an internal state space and dynamics that fail to reflect the true environment. The problem is that credit assignment (i.e., learning a mapping from working memory to actions) is only useful inasmuch as the internal state corresponds to the true hidden state of the POMDP, leading to a chicken-and-egg problem.

Given the preceding argument, one obvious modification that might lead to improved performance is to increase the number of memory elements. As the number of memory elements increases, the probability that the model remembers observation A for the required amount of time approaches one. However, this strategy introduces the curse of dimensionality due to the rapidly increasing size of the internal state space.

This intuitive analysis suggests a normative explanation for the famous "magic number" limitation observed in human working memory capacity, thought to be about four independent elements (e.g., Cowan, 2000). We demonstrate this idea by again simulating the artificial grammar task, this time averaging performance over a range of training times (1 to 10 million time steps) to capture the idea that humans may practice novel tasks for a typical, but variable, amount of time. Indeed the averaged

results show diminishing returns of increasing memory elements (Figure 2C). This simulation used tabular (rather than more neurally plausible) representations and a highly simplified model, so the exact number of policy parameters and state values to be estimated, time steps, and working memory elements is somewhat arbitrary in relation to human learning. Still, the model's qualitative behavior (evidenced by the shape of the resulting curve and the order of magnitude of the optimal number of working memory elements) is surprisingly reminiscent of human behavior. Based on this we suggest that the limitation on working memory capacity may be due to a limitation on learning rather than on storage: it may be impractical to learn to utilize more than a very small number (i.e., smaller than 10) of independent working memory elements, due to the curse of dimensionality.

## 2.4 Discussion

We have presented a psychological model that suggests that dopaminergic PE signals can implicitly shape working memory representations in PFC. Our model synthesizes recent advances in the Gating literature (O'Reilly & Frank, 2006) with normative RL theory regarding model-free, finite memory solutions to POMDPs (Peshkin et al., 1999). We showed that the model learns to behave optimally in the benchmark 12-AX task. We also related the model's computational limitations to known limitations of human working memory (Cleeremans & McClelland, 1991; Cowan, 2000).

### 2.4.1 Relation to Other Theoretical Work

Other recent work in neural RL has argued that the brain applies memory-based POMDP solution mechanisms to the real-world problems faced by animals (Daw, Courville, & Touretzky, 2006; Samejima & Doya, 2007; Yoshida & Ishii, 2006; Dayan, 2007). That work primarily considers model-based mechanisms, in which the tem-

porary memory is a continuous belief state, and assumes that a function of cerebral cortex is to learn the required world model, and specifically that PFC should represent temporary goal- or policy-related information necessary for optimal POMDP behavior. The model that we present here is related to that line of thinking, demonstrating a model-free, rather than model-based, mechanism for learning to store policy-related information in PFC. Different learning systems may form different types of working memory representations. Future work may investigate the relationship between implicit learning (as in this Gating model) and model-free POMDP solutions, versus other kinds of learning and model-based POMDP solutions. Irrespective of the POMDP framework, other work has assumed that there exists a gating policy that controls task-relevant working memory updating in PFC (e.g., Dayan, 2007). The present work further develops a model of how this policy can be learned.

It is interesting to compare our model to previous work on model-free POMDP solutions. McCallum (1995) first emphasized the importance of learning utile distinctions, or learning to resolve two hidden states only if they have different optimal actions. This is an emphasis that our model shares, at least in spirit. Humans must of course be extremely flexible in their behavior. Therefore there is an inherent tension between the need to focus cognitive resources on learning the immediate task, and the need to form a basis of general task knowledge (Rougier et al., 2005). It would be interesting for future work to explore how closely the working memory representations learned by our model align to McCallum's utile (and less generalizable) distinctions as opposed to more generalizable representations of the underlying hidden structure of the world, or whether our model could be modified to incorporate a mixture of both kinds of knowledge, depending on some exploration/exploitation parameter.

Our model most closely follows the Gating model described by O'Reilly and Frank (2006), and the theoretical model described by Peshkin et al. (1999). Our model is clearly more abstract and less biologically detailed than O'Reilly and Frank (2006).

However, our intent was to ask whether the important insights and capabilities of that model could be captured using a four-parameter, pure RL model with a clear normative basis. Accordingly, we have shown that such a model is comparably equipped to simulate a range of psychological phenomena. Our model also makes equally testable (albeit different) predictions about the neural DA signal. Relative to Peshkin et al. (1999), our model places biological and psychological concerns at the forefront, eliminating the episodic memory requirements of the Monte-Carlo algorithm. Considering Peshkin et al. (1999), it is perhaps interesting that our model performed so poorly when $\lambda = 1$, as this produces a nearly Monte-Carlo scheme. The difference was likely due to our model's online learning (i.e., we updated the policy at each time step rather than at the ends of episodes), which invalidates the Monte-Carlo approach. Thus it might be said that our model is a uniquely psychological variant of that previous architecture.

### 2.4.2 Implications for Working Memory and Cognitive Control

Subjects in cognitive control experiments typically face situations in which correct behavior is indeterminate given only the immediate observation. Working memory is often thought of as the repository of temporary information that augments the immediate observation to permit correct behavior, sometimes called goals, context, task set, or decision categories. These concepts are difficult to define. Here we have proposed a formal theoretical definition for the cognitive control and working memory constructs. Due to the importance of temporally distant goals and of information that is not immediately observable, the canonical cognitive control environment is well captured by a POMDP. Working memory is then the temporary information, defined and updated by a memory control policy, that the animal uses to solve these POMDPs. Model-based research might identify working memory with continuous

belief states, whereas our model-free framework identifies working memory with a discrete collection of recent observations. These may correspond to the products of different learning systems, but the outcome is the same in either case: cognitive control is defined as an animal's memory-based POMDP solver, and working memory is defined as the information, derived from recent history, that the solver requires.

### 2.4.3 Psychological & Neural Validity

Although the intractability of solving a POMDP means that all models such as the one we present here must ultimately fail to find an optimal solution in a practical amount of time (if at all), the particular manifestation of computational limitations in our model aligns qualitatively with that observed in humans. Working memory, the psychological construct that the Gating model addresses, is famously limited (see Cowan, 2000, for a review). Beyond canonical working memory capacity limitations, other work has shown subtler limitations arising in learning contexts (e.g., Cleeremans & McClelland, 1991). The results that we presented here are promising, but it remains for future work to more fully explore the relation between the failures exhibited by this model and those exhibited by humans.

In conclusion, we have shown that the Gating framework provides a connection between high level cognitive concepts such as working memory and cognitive control, systems neuroscience, and current neural RL theory. The framework's trial-and-error method for solving POMDPs gives rise to particular limitations that are reminiscent of observed psychological limits. It remains for future work to further investigate the model's ability to capture a range of specific psychological and neural phenomena. Our hope is that this link between working memory and POMDPs will be fruitful in generating new insights, and suggesting further experimental and theoretical work.

# Chapter 3

# Within-Trial Dynamics: Model and Behavioral Experiment

## 3.1 Introduction

### 3.1.1 Task Switching

Humans are steeply limited in their multitasking ability, and the task switching literature serves as an experimental window into this limitation. This literature probes the factors underlying the "switch cost," or increase in reaction time (RT) when switching between two simple tasks as compared to when performing either task in isolation. This was first reported by Jersild (1927), and Spector and Biederman (1976) later established important boundary conditions under which the switch cost occurs. The modern task switching literature truly began in the 1990s however, when Allport et al. (1994), Rogers and Monsell (1995), and Meiran (1996) demonstrated the existence of and established the standard paradigm for measuring the "residual switch cost," or RT impairment that remains even after long preparatory intervals. The present work focuses on this residual switch cost, and in particular adjudicates between two families of hypotheses, motivational versus structural, regarding its underlying factors.

One family of hypotheses of the residual switch cost presupposes that participants are fully motivated to perform in the experiments. This led to an emphasis on motivation-insensitive structural cognitive factors. For example, Allport et al. (1994) interpreted the residual switch cost as evidence of task set "inertia," or lingering activation of the previous task that passively recedes, leading post-switch performance to be initially sluggish. Similarly, Altmann and Gray (2008) interpreted the residual switch cost in terms of proactive interference in episodic memory from previous task sets. Others (Rogers & Monsell, 1995; Meiran, 1996) interpreted the residual switch cost in terms of an active preparatory process with some components that do not begin until after the first post-switch performance.

An alternative family of hypotheses, put forward by De Jong and colleagues (De Jong, Berendsen, & Cools, 1999; De Jong, 2000) and later Nieuwenhuis and Monsell (2002), questions whether participants are indeed fully motivated, explaining the residual switch cost in terms of motivational factors. De Jong (2000) argued that reliably preparing for task switches is "optional" and effortful, and that residual switch costs arise because participants are unwilling to expend the requisite effort. This was supported by an analysis showing that residual switch RT distributions are well fit by mixtures of fast, "prepared" distributions, and slow, "unprepared" distributions, and that the observed (i.e., estimated) proportion of prepared trials is affected by motivational manipulations. Nieuwenhuis and Monsell (2002) followed up on De Jong's work. They re-analyzed Rogers and Monsell's (1995) original residual switch cost data and performed a replication with additional monetary motivation for fast, accurate responses, effectively implementing a between-subjects motivational manipulation. They found that the old and new data were well fit by De Jong's prepared-unprepared mixture model, and that motivation increased the estimated proportion of prepared trials by 33%.

Interestingly however, there is a speed-accuracy tradeoff in Nieuwenhuis and Mon-

sell's (2002) data.[1] Participants in the high motivation condition were not only faster but were also less accurate than those in the low motivation condition (Nieuwenhuis & Monsell, 2002). In earlier work, De Jong (2000) had explicitly tested for motivational effects on accuracy, but found a null result. However, De Jong's study included just eight participants, and may therefore have been underpowered to detect the accuracy effect. The possibility of a speed-accuracy tradeoff in these data is intriguing, because it would undermine support for the motivation-sensitive preparation hypothesis. That is, it is possible that increasing motivation could lead to different speed-accuracy tradeoffs in switch as opposed to repeat trials, leading to motivation by trial type (i.e., switch versus repeat) interactions qualitatively similar to those that would be caused by motivation-sensitive preparatory processes. For example, if participants are paid for accurate responses made before a deadline in a motivation condition, then the optimal speed-accuracy tradeoff on switch trials might be different than that on repeat trials. In this case, even if participants in fact exhibit a motivation-insensitive RT increase for switch as opposed to repeat trials, there would be a reduction in RT switch costs (but also a decrease in switch accuracy) in the motivation condition.

The potential for multiple causes of differences in two dependent variables (both RT and accuracy) across different motivation and trial type conditions underscores the need for precise hypothesis articulation and testing. A natural aid in this situation is rational analysis (e.g., Anderson, 1990, 1991), a theoretical framework in which motivational factors (i.e., costs and constraints associated with an experiment) specify a mathematical optimization problem, the solution of which serves as a model of participant behavior. I use rational analysis here to formalize the competing motivational and structural hypotheses of the residual switch cost into two alternative models: one

[1]The effect of motivation on accuracy was not tested in Nieuwenhuis and Monsell's (2002) original article. I obtained the data from Sander Nieuwenhuis to perform the test: collapsing across trial type and across all RSI conditions that were included in both experiments, a two-sample t-test revealed a significant effect of motivation on accuracy, $t(20) = 2.223$, $p = 0.03$.

in which motivation effects preparatory processes as well as the speed-accuracy tradeoff, and one in which motivation effects the speed-accuracy tradeoff only. This allows us to ask whether the data require the motivation effects on the preparatory process, after "controlling" for the motivation effects on the speed-accuracy tradeoff that are clearly evident from Nieuwenhuis and Monsell's (2002) data. To anticipate, I will show that both models are able to capture the qualitative patterns of previous data. I then conduct two new experiments to better adjudicate between the models. The first experiment is similar to that conducted by Nieuwenhuis and Monsell (2002), but uses a parametric, within-subject motivation manipulation. The second experiment was provides a converging, model-free assessment. Both new experiments provide evidence for motivation-insensitive preparatory processes. This casts doubt on the motivational interpretations of the residual switch cost put forward by De Jong (2000) and later Nieuwenhuis and Monsell (2002), and supports instead the structural interpretations (e.g., Allport et al., 1994; Rogers & Monsell, 1995; Meiran, 1996; Logan & Bundesen, 2003; Altmann & Gray, 2008).

### 3.1.2 Motivational Effects in Task Switching

The motivational effects on preparatory processes that I include in my analysis were based on De Jong's (2000) original idea that residual switch costs reflect a strategic tradeoff between mental effort and performance. The question of how mental effort relates to behavior has been treated in literatures on judgment and decision making (Bettman, Johnson, & Payne, 1990; Johnson & Payne, 1985; Shah & Oppenheimer, 2008), economic decision theory (Wilcox, 1993, 1994), cognitive control (Botvinick et al., 2009; Kool, McGuire, Rosen, & Botvinick, 2010), and visual attention (Ballard, Hayhoe, Pook, & Rao, 1997; Droll & Hayhoe, 2007). I apply ideas from these literatures, together with standard, general cognitive control assumptions (Cohen, Dunbar, & McClelland, 1990; Miller & Cohen, 2001; Norman & Shallice, 1986; Schneider &

Shiffrin, 1977), to formulate the precise assumptions that are necessary for the rational analysis of task switching. Assumptions will be discussed and justified in detail below, but, in summary, they are: (1) A task must be active in working memory before it can be performed. (2) Only one task can be active at a time. (3) Participants choose when to initiate task preparation. (4) Preparing a task (i.e., moving it into an active state) takes time. (5) Both initiating task preparation and maintaining a task in an active state incur mental effort costs. (6) Participants are motivated to perform well. Together, these assumptions imply a critical tension between effort costs (preparing earlier means paying greater task maintenance costs) and performance costs (preparing later means slower RT). Analysis of this tension reveals that the optimal resolution leads to residual switch costs.

In the sections that follow, I discuss the assumptions of my analysis in more detail. I then show that both models derived from these assumptions exhibit several known properties of the residual switch cost (Rogers & Monsell, 1995; De Jong, 2000; Nieuwenhuis & Monsell, 2002; Dreisbach & Haider, 2006; Verbruggen, Liefooghe, Vandierendonck, & Demanet, 2007). Finally, I apply this framework to analyzing two new experiments in order to adjudicate between the hypotheses under consideration.

## 3.2 The Model

### 3.2.1 Task Switching Experiment Structure

In a typical task switching experiment (Figure 3.1; see also Meiran, 1996), participants respond to a sequence of hundreds of trials. Each trial begins with a cue indicating the required task. On switch trials, this is different from the previous cue. On repeat trials, it is the same. The cue is followed by a CSI, generally on the order of one second, after which the target appears. Participants' goal is to respond to the target according to the cued task. Participants take a certain reaction time (RT) to respond,

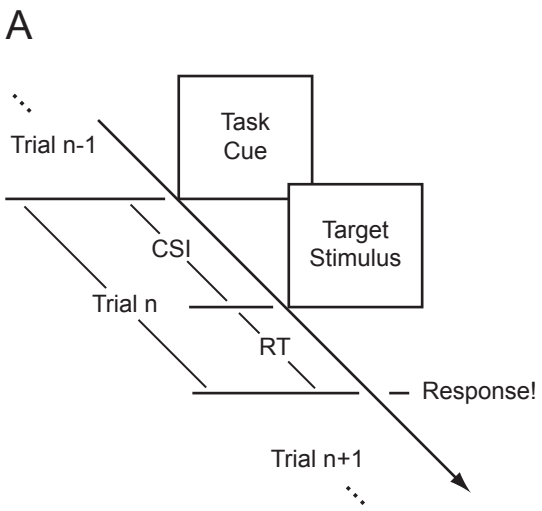



Figure 3.1: Task switching experiment trial structure.

after which the next trial begins. My model does not differentiate between exogenous and endogenous task cues, the latter coming from predictable (non-cued) switching schedules, and therefore I refer to the preparatory interval in all experiments as a CSI.

The target stimulus is bivalent, meaning that it is a suitable input to both tasks — Spector and Biederman (1976) established this as a boundary condition for switch costs to occur. Moreover, some or all targets are incongruent, meaning that each task prescribes a different response. To take a common parity/magnitude example, the parity task maps odd and even digits to left and right button press responses, respectively. In contrast, the magnitude task maps digits less than five and digits greater than five to left and right responses, respectively. In this example, the target '7' is a suitable input for either task (i.e., bivalent), but maps to a left response under the parity task, and a right response under the magnitude task (i.e., incongruent).

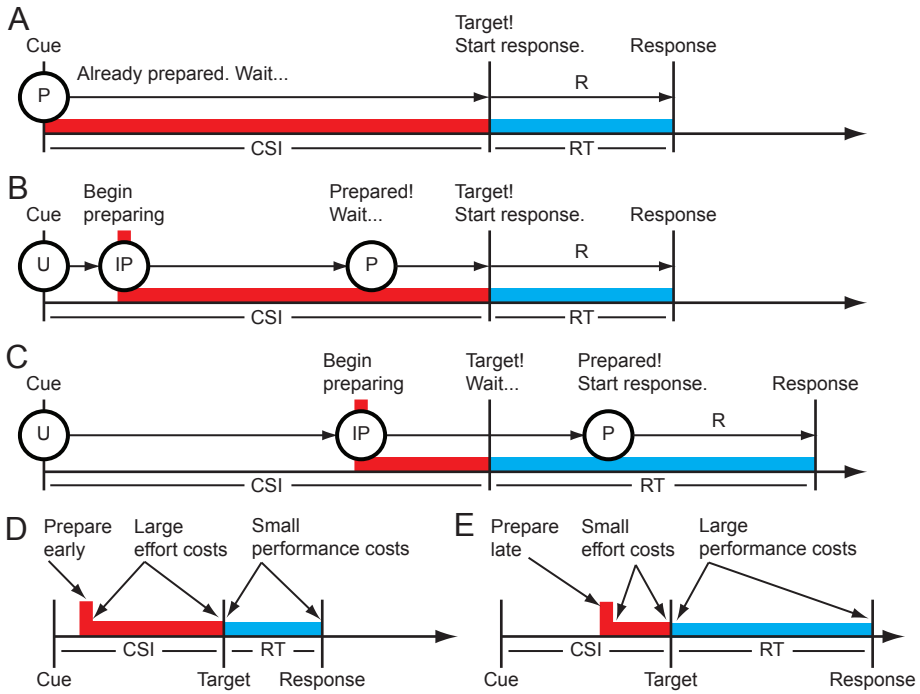



Figure 3.2: Rational analysis of task switching. Trial structure is described in text. Circles indicate mental state (U=unprepared, IP=in preparation, P=prepared). Red indicates mental effort costs. Cyan indicates performance costs. Participants must be prepared for the cued task prior to beginning the response process. Preparation takes time to complete, and I assume that participants pay effort costs for initiating as well as maintaining preparation. These must be weighed against performance pressures that translate into costs for long RTs. On repeat trials (A), participants are already prepared for the cued task, and begin responding immediately after target onset, minimizing performance costs. They pay maintenance costs for the entire CSI, but completely avoid the large preparation initiation cost. On switch trials, participants must choose when to prepare the new task. If participants begin preparation early (B), they pay greater maintenance costs, but preparation is likely to complete before target onset, minimizing effort costs. If participants begin preparation late (C), they pay lower maintenance costs, but preparation is unlikely to complete prior to target onset, leading to greater performance costs. Note that the response process duration (R) is the same in (B) and (C), but the total RT is longer in (C) because the response process does not begin until preparation completes. I apply rational analysis to this putative trade-off (D,E) between effort and performance costs.

### 3.2.2 Rational Analysis

To apply rational analysis to task switching experiments, one must interpret the experimental structure described above in terms of an optimization problem, including constraints, costs, and choices that participants face. First, because targets are bivalent and incongruent, participants can only be prepared for one task at a time. If they attempted to prepare for both tasks simultaneously, then they would be unable to respond reliably to targets, as the two tasks would prescribe incompatible motor responses. For simplicity, I assume that preparedness is binary for each task. A task is either prepared, driving the corresponding response, or it isn't, and the corresponding response cannot yet be made.

Second, I assume that mentally preparing a task takes a non-trivial amount of time. This follows from the interpretation of the empirical evidence (e.g., Rogers & Monsell, 1995; Meiran, 1996; De Jong, 2000; Monsell, 2003; Nieuwenhuis & Monsell, 2002) that assumes that the residual switch cost as evidence for the duration of an active preparatory process. Note that even the motivational accounts assume that the preparatory process takes time, differing from the structural accounts in the assumption that participants can choose when to initiate this time consuming process. Given my binary preparedness assumption, this implies that, once participants initiate preparation, the task does not enter its prepared state until the end of the preparatory process. That is, during the time that the task is "in preparation," participants are not yet able to begin the response process.

Moreover, I assume that there is stochasticity or randomness associated with the preparation process. This follows from empirical evidence specific to task switching (e.g., De Jong, 2000; Nieuwenhuis & Monsell, 2002), and is a widely used assumption regarding cognitive processes in general (e.g., memory retrieval times: Ratcliff, 1978). Specifically, I assume that, once initiated, preparation takes a random time to complete. For simplicity, I assume that this random time is exponentially distributed.

This random preparation duration assumption is similar to De Jong's (2000) Failure to Engage hypothesis. I offer a more thorough comparison below.

Third, I assume that participants experience aversive mental effort costs associated with task preparedness. I assume that initiating preparation leads to a transient, large cost, perhaps due to a general disruption of working memory following the updating of a particular piece of information, or to a metabolically expensive update process. I further assume that, once preparation is initiated, there is an ongoing maintenance cost per unit time, because of the opportunity costs and physiological requirements of dedicating part of working memory to perform the task.

Fourth, I assume that participants experience performance costs in task switching experiments. In a minority of task switching experiments (e.g., Nieuwenhuis & Monsell, 2002; Savine & Braver, 2010), participants actually receive performance-based monetary incentives. However, in almost all task switching experiments, experiment instructions (the ubiquitous instruction to respond as quickly and accurately as possible), demand characteristics (e.g., the desire to impress the experimenter) and opportunity costs (i.e., the desire to finish the experiment as quickly as possible) can be interpreted as implicit performance costs. I interpret all of these performance pressures as one cost for incorrect responses, and another cost per unit of RT.

The preceding assumptions are sufficient to analyze RT qualitatively on both switch and repeat trials. This is shown in Figure 3.2. I focus on the first moment of the CSI, when the participant has perceived and interpreted the cue. On a repeat trial (Figure 3.2A), the participant is already prepared for the cued task. The participant may choose to forget the task and re-prepare it at a later point in the CSI, or to simply maintain the task throughout the CSI. Although maintaining the task throughout the CSI incurs greater maintenance effort costs, this strategy avoids the large preparation initiation cost. Assuming that the transient cost is large enough to warrant this strategy, consider the ensuing performance costs. They are clearly minimized, because

the task is guaranteed to be prepared prior to target onset, causing the response process to begin immediately following target onset, leading to the fastest possible RT for a given level of accuracy.

Now consider the case of a switch trial (Figure 3.2B-C). Here the participant is initially prepared for the wrong task, and it is obviously rational to stop maintaining the now inappropriate task immediately in order to avoid effort costs. The participant now faces the dilemma of when to initiate preparation for the new task. Preparing early leads to large maintenance effort costs but a high probability of being fully prepared prior to target onset and thus minimizing performance costs (i.e., minimizing RT). Preparing very late leads to small maintenance effort costs but a low probability of being fully prepared prior to target onset and thus to large performance costs (i.e., a long RT).

Up to now, I have not assumed anything regarding the response process. However, this is an important piece of the analysis, because it is here that the speed-accuracy tradeoff applies. I assume a drift-diffusion model (DDM: e.g., Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Gold, 2002; Ratcliff, 1978) for the response process. This is a standard way of modeling responses that are driven by a speed-accuracy tradeoff that is optimized by the participant. This is described in more detail in section 4.5.

The assumptions described above are parametric. There are eight free parameters in total, including two mental effort costs, two performance costs, a rate for the preparation process, and three DDM response parameters. These parameters are fit to various datasets as described below.

I have now specified all of the constraints and costs that define the rational analysis of task switching. The central idea of the rational analysis framework is that participants make choices that are optimal given these costs and constraints. Choices can be optimized separately in different experiment conditions, inducing a pattern of

optimal behavior across conditions that forms a model of data. Optimal choices are derived in the Appendix.

I define the two different models by manipulating the set of choices that are available to participants. In both models, participants choose repeat trial strategy (i.e., maintain or forget and re-prepare) and speed-accuracy tradeoff (in the DDM response process). In the motivation-sensitive preparation model, participants also choose preparation initiation time (somewhere between the first instant of the CSI and the time of target onset). In the motivation-insensitive preparation model, participants do not choose preparation initiation time. Instead, the asynchrony between target stimulus onset and preparation initiation time is a free parameter of that model, and is generally fixed across experiment conditions.

To conclude, the rational analysis of task switching includes a few fundamental assumptions. First, task switching participants must be prepared for a task before performing it, and can only be prepared for one task at a time, leading to a problem of when to prepare the appropriate task on each switch trial. Second, task preparation takes a non-trivial amount of time, which is assumed to be exponentially distributed. Third, participants experience mental effort costs both for initiating and maintaining task preparedness. Fourth, participants face pressure to respond as quickly and accurately as possible, which translates into performance costs for error responses and per unit RT. Fifth, participants control the basic response process by setting the speed-accuracy tradeoff, which is expressed specifically with a DDM. Finally, participants make choices that minimize costs given the constraints. The set of which choices participants are allowed to make is used to differentiate the two competing models under consideration. In the next section, I examine how these assumptions give rise to optimal behavior that is similar to data from previous task switching experiments.

### 3.2.3 Exploring Model Behavior in Pre-Existing Experiments

In the sections that follow, I explore how my rational analysis of task switching can reproduce phenomena from previous experiments. Here I consider the qualitative patterns of mean accuracy and RT across different experiment conditions. The motivation-sensitive preparation model and the motivation-insensitive preparation model yield essentially identical fits to these mean patterns, and I present the motivation-sensitive preparation model results as a proxy for both models. I fit model parameters by minimizing sum of squared errors (SSE). That is, I find model parameters such that optimal mean RT and accuracy are as close as possible to published data in the squared error sense. To accommodate overall differences in mean RT and accuracy across experiments, separate drift rate parameters (for the DDM response process) were fit for each experiment. All other model parameters are constrained to be identical across experiments. It should be noted that this procedure corresponds to regressing just 28 data points (the conditional means of previous experiments, described below) onto a 10-parameter nonlinear model. This is too little data to make reliable inferences about fitted parameter values, and I intend it simply as a principled way of setting parameters in order to illustrate model properties.

#### Residual Switch Costs (Rogers & Monsell, 1995)

Rogers and Monsell found in their seminal Experiment 3 (1995) that switch costs decreased as the experimenters increased the CSI — yet the rate of decrease leveled out when the CSI was around 600ms, leaving robust RT *residual switch costs*. In contrast, they found that error costs were only robust for small CSIs, with no error residual switch costs. Here I explore the rational analysis model's behavior in the context of this experiment.

First, note that preparation cannot begin before the start of the CSI, as this would lead to conflicting responses to incongruent targets. Therefore, although the

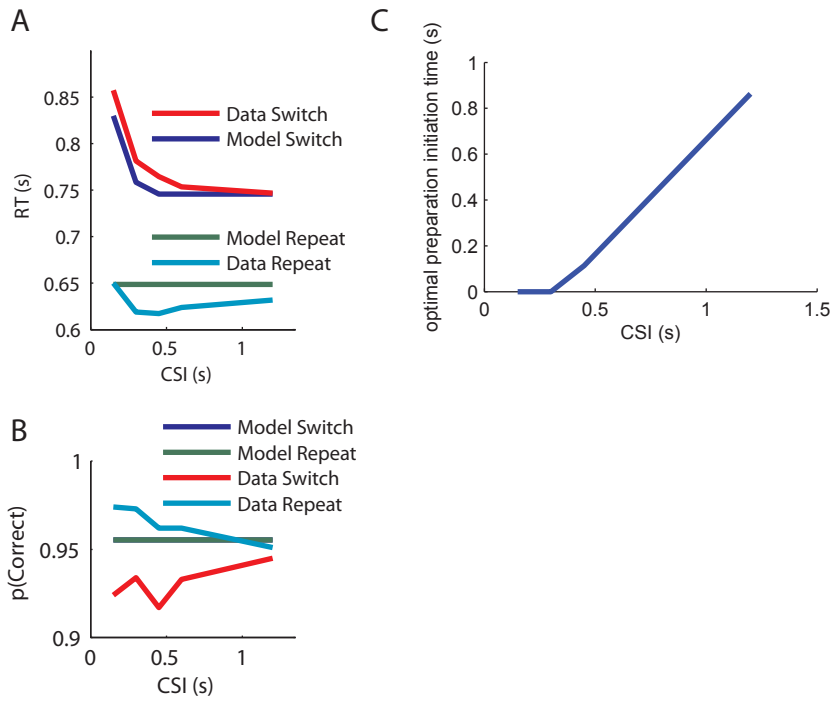



Figure 3.3: Model analysis and data from Rogers & Monsell's (1995) Experiment 3. (A) The model exhibits RT residual switch costs. (B) Neither CSI nor switching affects the model's error rate. (C) Shifting CSI causes an identical shift in the cost-by-preparation initiation time curve (solid lines), and in the optimal preparation initiation time (dashed lines). Actual preparation initiation time cannot be negative. (D) As the maintenance cost goes to zero, optimal preparation initiation time accelerates toward negative infinity. Dashed line indicates fitted maintenance cost.

*optimal* preparation initiation time may be negative for some CSIs, *actual* preparation initiation time is necessarily zero or positive. Next, I simply compute the optimal preparation initiation time (or zero, for negative optimal times) for each CSI condition. From this, I determine the model's mean RT and accuracy (derivations in the Appendix).

Figure 3.3A-B shows the original data together with the model's behavior in this setting. The model exhibits the same qualitative RT pattern as the data, with a switch-repeat difference that starts out large, but then decreases as CSI increases, until leveling out around 600ms. To understand the pattern of decreasing switch trial RT across CSIs, note that optimal preparation initiation time is simply a constant plus the length of the CSI. In the motivation-insensitive preparation model, preparation time is not optimized but is still a constant (the asynchrony parameter) plus the CSI. Thus, in both models, as CSI decreases, preparation initiation time decreases by the same amount. However, since actual preparation initiation time cannot be negative, decreasing CSI ultimately "squashes" preparation initiation time against zero. This effectively makes preparation initiation time *later*, leading to a greater probability that task preparation will complete after target onset, which leads in turn to a delayed response process and longer RT. Figure 3.3C shows how the curve relating expected total costs (i.e., effort costs plus performance costs) to preparation initiation time shifts to the left as CSI decreases.

The model also produces repeat trial RTs that are constant across CSIs. This is because the effort cost associated with preparation initiation is large enough that the optimal repeat trial strategy is to maintain task preparation throughout the CSI. This leads the model to be always fully prepared prior to target onset, leading to minimum RTs, irrespective of CSI.

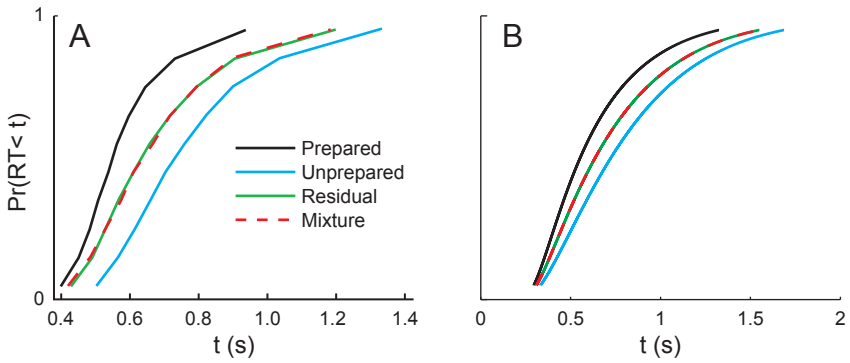



Figure 3.4: Failure to Engage hypothesis. (A) Data from Rogers & Monsell (1995), as re-analyzed by Nieuwenhuis & Monsell (2002). The long-CSI switch distribution (green line) is a mixture (red dashed line) of long-CSI repeat (black line) and short-CSI switch (cyan line) distributions, as shown by the overlapping red and green lines. (B) The model exhibits the mixture pattern, with the red and green lines again overlapping.

### Failure to Engage (De Jong, 2000)

De Jong (2000) argued that preparation on long-CSI switch trials is "optional," and that participants prepare on some trials but not on others. He supported this argument by claiming that the distribution of RTs in long-CSI switch trials — i.e., the distribution of RT residual switch costs — is a mixture of a fast "prepared" distribution and a slow "unprepared" distribution. Specifically, he claimed that prepared trials were statistically equivalent to long-CSI repeat trials, and that unprepared trials were equivalent to short-CSI switch trials. This idea later received some support from Nieuwenhuis and Monsell (2002), in that they re-analyzed Rogers & Monsell's (1995) data and found that the mixture model fit the data well. Figure 3.4A shows Nieuwenhuis & Monsell's re-analysis of Rogers & Monsell's (1995) data in terms of this mixture hypothesis.

The model exhibits this mixture property in most experimental settings, because of the exponential random duration of the preparation process. That is, on repeat trials, assuming that it is optimal to maintain preparation throughout the CSI, the

participant is fully prepared prior to target onset, leading to an RT distribution that is simply that of the DDM response process. On switch trials, the participant may or may not be fully prepared prior to target onset. On fully prepared switch trials, the RT will again be a simple DDM response process, making it indistinguishable from repeat trial RTs. On unprepared switch trials, the response process does not begin until preparation completes. In this case, the observed RT is the sum of the remaining preparatory process duration and the DDM response process. This remaining preparatory process duration has a special property that follows from the assumption that the preparatory process duration is exponentially distributed. Specifically, given that the preparatory process has not completed prior to target onset, the remaining preparation time is statistically the same as if preparation had just begun. This follows from the "memoryless" property of the exponential distribution. Given this property, all unprepared trials have the same RT distribution, regardless of CSI. Thus, any particular long-CSI switch RT distribution does indeed come from the same distribution as either the short-CSI switch RT or long-CSI repeat RT, as argued by De Jong (2000).

Although this memoryless property is unique to the exponential distribution, an alternate assumption about preparatory process randomness would also produce the mixture pattern that is observed in the data. The alternate assumption is that, rather than participants picking the deterministic initiation time of a random-duration preparatory process, they pick the probability that a fixed-duration preparation process will begin at cue onset rather than target onset. This alternate assumption is perhaps closer to De Jong's (2000) original Failure to Engage hypothesis (although the original hypothesis was not specified at this level of detail), and would lead to similar behavior of my overall rational analysis model. However, I chose the random-duration model, with the exponential distribution, because the fixed-duration assumption leads to the strong prediction that the short-CSI switch RT distribution is simply a right-

shifted version of the long-CSI repeat distribution. In the data, these distributions appear to have different shapes, with the short-CSI switch RT distribution exhibiting a heavier tail. The random-duration assumption leads to this heavier tail.

Finally, the model distributions generally have heavier tails than the corresponding data distributions. This is due in part to the fact that the model was fit only to the conditional means of previous experiments, rather than to the entire distributions. However, another issue is that the standard DDM response model assumes that the response process proceeds based on a constant rate of evidence accumulation, and that the process terminates only when evidence has reached a certain threshold, irrespective of the duration of the ongoing response. In contrast, real participants may have a collapsing decision threshold, reflecting response impatience that increases with each additional millisecond. Real response processes are likely to be far more complicated, including growing impatience and impulsivity as the response duration drags on. For these two reasons, I make no strong claims about distribution shape in this section, other than to illustrate that the model reproduces the mixture property.

### Probabilistic Switch Cues (Dreisbach & Haider, 2006)

Dreisbach and Haider (2006) used probabilistic switch cues, allowing them to examine the effect of uncertainty about the upcoming task on switch costs. At the beginning of each trial, the cue indicated only the probability that the task would repeat (75% vs 25%). At the time of target onset, an additional cue (the color of the target) revealed the actual task. When the probability of a repeat trial was 75%, they found typical RT residual switch costs. In contrast, when the probability of a repeat trial was just 25%, switch costs disappeared due to the fact that repeat trial RT became as slow as switch trial RT (Figure 3.5A).

Both models reproduce this pattern (Figure 3.5B). Rational analysis of this experiment hinges on the fact that participants are faced with a strategic choice at the

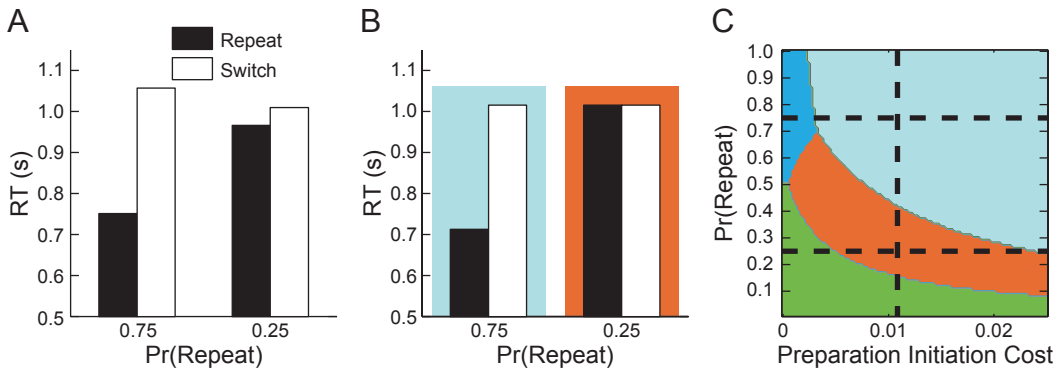



Figure 3.5: Probabilistic switch cues. (A) Data from Dreisbach & Haider (2006). (B) Model behavior. (C) Optimal strategy (shading) as function of probability of repeating and preparation initiation cost. Light blue=maintain currently prepared task in anticipation of a repeat. Green=forget currently prepared task and prepare for a switch. Dark blue=forget currently prepared task and re-prepare for a repeat. Orange=forget currently prepared task and do not prepare until the true task is revealed at target onset. Horizontal lines denote switch probabilities used in the experiment. Vertical line denotes preparation initiation cost (arbitrary units) estimated from previous experiments.

beginning of each trial: they can maintain the currently prepared task in anticipation of a repeat trial; forget the currently prepared task and not prepare any task until the true task is revealed at target stimulus onset; or forget the currently prepared task and prepare for a switch. The choice of strategy drives RT (and performance costs). This applies to both of my models, as the motivation-insensitive preparation model retains the choice of whether or not to maintain information or forget and re-prepare. For example, if participants choose the maintenance strategy, then they will be prepared to respond immediately upon target onset if the true task is in fact a repeat. In case of a switch however, they will be completely unprepared and will have to initiate preparation at the time of target stimulus onset, leading to maximum RT. Mental effort costs are also an important factor in strategy choice. To continue this example, in the case of a repeat, the participant has successfully avoided paying the large preparation initiation cost. However, in case of a switch, the participant has paid not only the preparation initiation cost at the time of target onset, but also the

maintenance cost for the wrong task throughout the entire CSI. Thus, the particular constellation of effort costs, performance costs, and switch probability drives strategy choice (and with it, RT), in a straightforward way.

Figure 3.5C illustrates how the optimal strategy changes as a function of probability of repeating and and the preparation initiation effort cost. The figure shows that, given the estimated preparation initiation cost (vertical bar), a 25% chance of repeating is not enough to warrant paying the maintenance effort cost throughout the CSI, but it is also not enough of a chance of a switch to warrant preparing for a switch (which would incur the risk of paying the preparation initiation cost twice). Thus, the optimal strategy is to forget the currently prepared task and not prepare for any task until the true task is revealed at the time of target onset. In contrast, a 75% chance of repeating is enough to warrant paying the maintenance cost and avoiding the initiation cost, in anticipation of repeating.

### 3.2.4 Model Summary

In this section, I presented the details of the rational analysis of task switching experiments. I also showed that, under these assumptions, a rational agent will produce behavioral patterns similar to those that have been reported in the task switching literature. Both the motivation-sensitive preparation model and the motivation-insensitive preparation model fit these previous data equally well. In the following section, I test the two competing models with newly collected data from a deadlined task switching paradigm.

## 3.3 Testing the Models: Deadlined Task Switching Paradigm

In the following sections, I describe new experiments used to select between the competing motivation-sensitive and motivation-insensitive models. To accomplish this, I employ a "deadlined" task switching paradigm that allows direct manipulation of performance costs. In this paradigm, participants earn monetary reward for correct responses made before a deadline. Correct responses made after the deadline earn nothing, and error responses made at any time incur a monetary penalty. This permits explicit manipulation performance costs, whereas when considering canonical task switching experiments, performance costs were assumed post hoc and were either implicit (and constant) in the original experiment (e.g., Rogers & Monsell, 1995), or were categorical rather than parametric, and difficult to associate with precise monetary values (e.g., Nieuwenhuis & Monsell, 2002). I designed my experiments in order to address these issues and yield datasets that would cleanly discriminate the two models.

### 3.3.1 Experiment 1

In Experiment 1, I used a parametric, within-subject motivational manipulation that would allow us to quantitatively select between the two competing models with a dataset that was conceptually similar to Nieuwenhuis and Monsell's (2002) experiment. I manipulated response deadline parametrically across blocks. In terms of the rational analysis, this manipulation effectively changed the performance cost, with changing deadlines rendering fixed RT and accuracy distributions more or less likely to yield monetary rewards. Both models predict an overall U-shaped trend as deadline increases from very short to very long, both in RT and accuracy. The U shape arises because, when the deadline is very short there is little chance of earning the

reward, and when the deadline is very long there is little chance of missing the reward. In either case, there is no reason to spend mental effort by preparing early, or to risk making an error by lowering the DDM threshold. For intermediate deadlines however, effort expenditure should be at a maximum and DDM threshold should be lowered in order to respond quickly when it is just barely possible to beat the deadline. Moreover, this pattern of motivational effects interacts with trial type, so that it is stronger on the (overall slower) switch trials, leading to a predicted trial type by deadline interaction. In both models, the changing DDM threshold contributes to the U-shaped main effect as well as to the U-shaped interaction. Thus the predictions of both models are qualitatively similar, which is why it was difficult to adjudicate between them based on Nieuwenhuis and Monsell's (2002) work in the first place. I therefore rely on quantitative model selection to select between the two models for Experiment 1.

### Methods

**Participants** 31 participants (20 female; mean age 20.4 years) were recruited from the Princeton University community. Participants were paid a base rate of $12 for the 50 minute experiment, plus a performance-based bonus amount that was typically around $10. Although bonuses could be negative (i.e., in the case of unusually poor performance), participants could not earn less than $12 for the session. The experiment was approved by the Princeton University human subjects research ethics board.

**Materials & Apparatus** Stimuli were presented and responses were recorded using Apple iMac computers with 24 displays, running Matlab (The MathWorks Inc., Nattick, MA) with Psychophysics Toolbox (Brainard, 1997). Responses were made using the '<' and '>' keys on a standard 101 key keyboard.

**Procedure** Participants performed parity/magnitude task switching. That is, on

each trial they were presented with a single digit, excluding five, and were required to judge whether it was odd or even (parity task), or greater or less than five (magnitude task), depending on the trial. "Odd" and "less than five" mapped to the '<' response key; even and greater than five mapped to the '>' response key. Participants were given instructions and practice, to ensure that they understood the experiment. They then progressed through 8 experimental blocks of cued (i.e., unpredictable order) task switching, separated by self-paced breaks with performance feedback.

Within each experimental block, deadline was fixed at one of four levels (0ms, 600ms, 1,000ms, 1,400ms). Deadline order was randomized across participants, but constrained so that, within participant, deadline order was the same across the first four blocks as across the four blocks. Each block contained 187 trials. Each trial began with a 900ms CSI. After each trial, the task always had a 0.5 probability of switching. Switch trials were indicated by a task cue that was presented throughout the CSI. The target appeared at the end of the CSI. After the participant responded to the target, a "reward thermometer" appeared for 200ms to indicate graphically whether the participant had earned or lost any money for the trial. Then the next trial began.

The reward for correct, on time responses, and the penalty for error responses were fixed at \$0.01 and -\$0.01, respectively. That is, for a correct response made before the deadline, participants earned \$0.01. For a correct response made after the deadline, participants earned \$0. For an incorrect response made at any time, participants lost \$0.01.

### Results

Experiment results are shown in Figure 3.6. ANOVA-based analyses are presented first, followed by quantitative model selection below. For RT, a trial type (Switch vs Repeat) by deadline (0ms, 600ms, 1,000ms, 1,400ms) repeated measures ANOVA

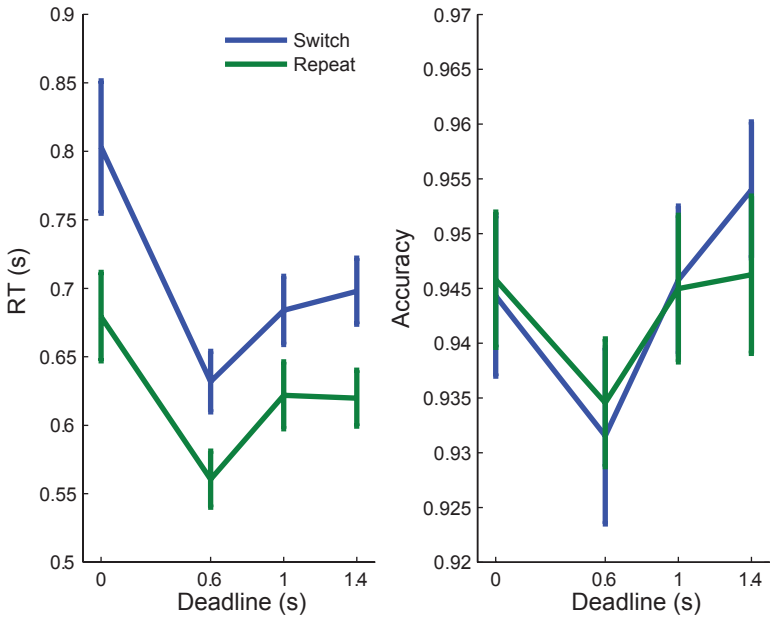



Figure 3.6: Experiment 1 RT and Accuracy.

revealed significant main effects of switch, $F(1, 30) = 24.36$, $p < 0.001$, and deadline, $F(3, 90) = 11.74$, $p < 0.001$. The predicted interaction was also significant, $F(3, 90) = 4.94$, $p = 0.003$. Given the qualitatively U-shaped model predictions and the approximately regular progression of the levels of my deadline manipulation, I tested quadratic trends using polynomial contrasts. The main quadratic trend was significant, $F(1, 30) = 13.69$, $p < 0.001$, as was the interaction, $F(1, 30) = 12.30$, $p < 0.001$, indicating that the quadratic trend across deadline levels was significantly stronger for switch than for repeat trials.

For accuracy, an identical ANOVA revealed a main effect of deadline, $F(3, 90) = 5.56$, $p = 0.002$, and a significant quadratic trend over deadlines, $F(1, 30) = 6.32$, $p = 0.017$. No other effects were significant.

To conduct formal model selection, each model was fit separately to each participant's data via maximum likelihood estimation (MLE), and then AIC and BIC were both used to compare model fits. BIC selected the motivation-insensitive preparation model in 21 of 31 participants. AIC selected the motivation-insensitive preparation

model in 25 of 31 participants. A binomial test rejected the null hypothesis that BIC (AIC) favors both model equally in the population of participants, $p < 0.05$ ($p < 0.001$).

#### Experiment 1 Discussion

Experiment 1 replicated and extended the findings of De Jong (2000) and Nieuwenhuis and Monsell (2002), showing that, with a parametric, within-subject motivational manipulation, RT (and residual switch cost) does indeed follow the predictions of a rational analysis account. Specifically, there is a U-shaped curve as response deadline increases, with a greater U-shape in switch than in repeat trials, as predicted by both models derived from my rational analysis. However, as I predicted, similar effects were also observed in accuracy, requiring consideration of a speed-accuracy tradeoff. Quantitative model selection results favored the motivation-insensitive preparation model over the motivation-sensitive preparation model. That is, my results support a reinterpretation of De Jong's (2000) and Nieuwenhuis and Monsell's (2002) data in terms of a speed-accuracy tradeoff, without motivational effects on preparation. Specifically, it seems likely that an interaction between speed-accuracy tradeoff and trial type led to nuanced effects that partially mimicked a motivation-sensitive preparatory process.

The design of Experiment 1 allowed us to reinterpret previous work on the strength of my own results. However, the design did not afford qualitative separation between the predictions of the two competing models. That is the aim of Experiment 2.

### 3.3.2 Experiment 2

In order to qualitatively separate the predictions of the two competing models, I drew upon the work of Ballard et al. (1997), which suggests that participants minimize mental effort by looking at visually presented information when they wish to store

it in working memory. Applying this logic to task cues, I assumed that participants would choose to look at the task cues at the time of task preparation initiation. Under this assumption, if the U-shaped curves that I observed in Experiment 1 were due in part to changes in preparation initiation time, then there should be corresponding U-shaped changes in "cue peek" times across deadline conditions. In contrast, if the U-shaped curves were due entirely to changes in the speed-accuracy tradeoff, there should be no changes in cue peek time across deadline conditions.

To test this, I used an experiment design that was similar to that of Experiment 1, except that a third task, digit color naming, was included, and on switch trials, participants were required to choose when to "peek" at the task cue. That is, on switch trials, the CSI began with an ambiguous switch cue indicating that the previous task was no longer appropriate. Participants then had to choose when to press the spacebar in order to see a 150ms presentation of the cue for the specific upcoming task. Reward deadline was then manipulated across blocks, as in Experiment 1.

### Methods

**Participants** 20 new participants (14 female, mean age 20.8 years) were recruited from the Princeton University community using the Princeton Paid Experiments website. Participants were paid a base rate of $12 for the 50 minute experiment, plus a performance-based bonus amount that was typically around $10. Although bonuses could be negative, participants could not earn less than $12 for the session. The experiment was approved by the Princeton University human subjects research ethics board.

**Materials & Apparatus** Materials and apparatus were identical to Experiment 1.

**Procedure** Experiment 2 was similar to Experiment 1, except that it included a third task, digit color naming. Digits were either black or white. At the beginning

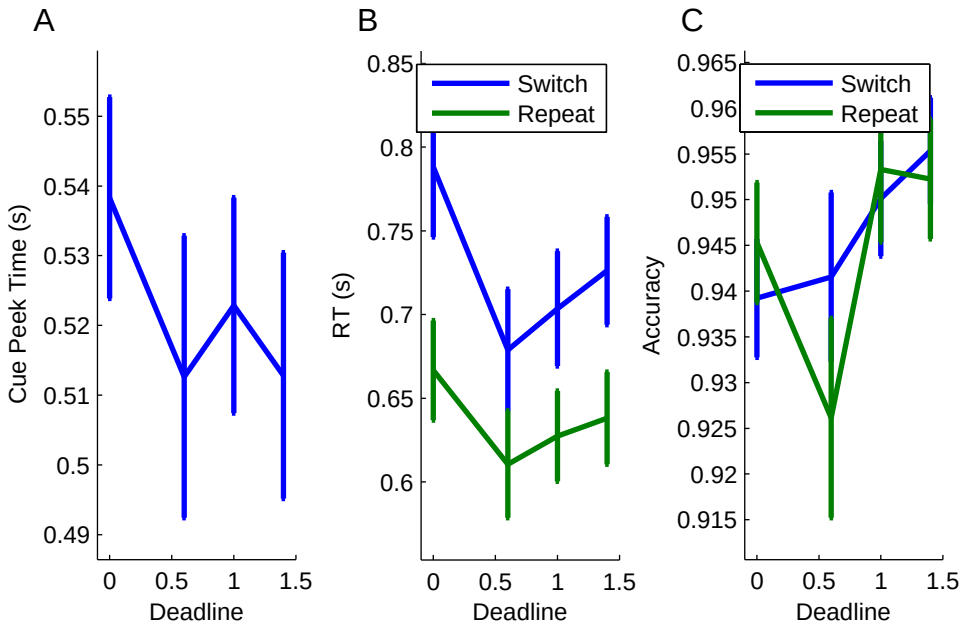



Figure 3.7: Experiment 2 Cue Peek, RT and Accuracy.

of repeat trials, participants saw a grey circle indicating that the required task was unchanged from the previous trial. At the beginning of switch trials, participants saw a yellow circle indicating that the required task had changed from the previous trial. Due to the fact that there were three tasks, this cue was only partially informative. Participants were required to press the spacebar at some point during the ensuing CSI in order to see the specific task cue for 150ms. Cue peek time, which was collected only on switch trials, was then an additional metric.

### Results

Experiment 2 results are shown in Figure 3.7. For RT and accuracy, results largely replicated those of Experiment 1. For RT, a trial type (Switch vs Repeat) by deadline (0ms, 600ms, 1,000ms, 1,400ms) repeated measures ANOVA revealed significant main effects of switch, $F(1, 19) = 21.87$, $p = 0.0002$, and deadline, $F(3, 57) = 5.01$, $p = 0.0038$. The trial type by deadline interaction was also significant, $F(3, 57) = 4.6$, $p = 0.0059$. For accuracy, there was a main effect of deadline, $F(3, 57) = 5.16$,

$p = 0.0032$, but no main effect of trial type, $p = 0.4836$. There was a significant trial type by deadline interaction, $F(4, 57) = 6.49$, $p = 0.0007$. Cue peek times could be collected only on switch trials, so that there was no trial type manipulation for this measure. For cue peek times, an ANOVA with factor deadline failed to detect a significant effect, $F(3, 57) = 1.38$, $p = 0.257$.

#### Experiment 2 Discussion

Experiment 2 included an additional metric of task preparation initiation time, and provided converging evidence, with Experiment 1, for the motivation-insensitive preparation interpretation of residual switch costs. That is, despite the fact that there was again a U-shaped pattern of RT and accuracy effects across increasing reward deadlines, this effect was not reflected in the time at which participants chose to "peek" at the task cue. Thus, whatever process drove the U-shaped RT and accuracy responses, it did not appear to relate to changes in the time of preparation initiation.

## 3.4 General Discussion

By appealing only to a straightforward interpretation of the costs, constraints, and choices implied by task switching experiments, I developed a rational analysis of task switching experiments that permits the precise articulation of competing hypotheses about motivational effects in this domain. In particular, the framework formalizes De Jong's (2000) ideas about how an effort-performance tradeoff e.g., Ballard et al., 1997; Johnson and Payne, 1985; Wilcox, 1993 might drive RT and accuracy in task switching. Critically, my framework goes beyond previous work by including a motivation-sensitive speed-accuracy tradeoff (e.g., Bogacz et al., 2006) in order to control for this source of variance when interpreting results. I demonstrated the overall plausibility of this framework in terms of its ability to explain existing data, including

residual switch costs (e.g., Rogers & Monsell, 1995), the failure-to-engage pattern (De Jong, 2000; Nieuwenhuis & Monsell, 2002), and the effects of probabilistic task cueing (Dreisbach & Haider, 2006). The framework is normative, in that it derives from the putative constraints imposed by the experiment, but also mechanistic, in that it prescribes particular sequential dynamics for a well known response model, the DDM.

I then used this framework to specify two competing models — a motivation-sensitive preparation model consistent with the work of De Jong (2000), Nieuwenhuis and Monsell (2002) and a motivation-insensitive preparation model consistent with the work of Allport et al. (1994), Rogers and Monsell (1995), Meiran (1996) — and I conducted new experiments to adjudicate between these. The experiments provided converging evidence in support of the motivation-insensitive preparation model. Experiment 1 used a deadline manipulation and was conceptually similar to the experiments reported by De Jong (2000) and Nieuwenhuis and Monsell (2002), except that it was a parametric, within-subject design. This replicated and extended the previous findings, validating the general prediction that RT and residual switch costs should respond in a U-shaped pattern to increasing response deadlines. However, model-based analysis showed that the data were better explained by the motivation-insensitive preparation model than the motivation-sensitive preparation model, in disagreement with De Jong's and Nieuwenhuis and Monsell's earlier interpretations of this type of data pattern. Experiment 2 sought a more direct measure off task preparation initiation time, applying the logic from Ballard et al. (1997) to assume that participants would choose to look at the task cue only at the time of preparation initiation in order to minimize mental effort. I reasoned that, if preparation initiation time was contributing to the U-shaped effect of deadline on residual switch costs, this U shape should also be reflected in cue peek times. However, despite the fact that RT and accuracy again exhibited U-shaped deadline effects, I found no effect of deadline

on cue peek times. This suggested that the process driving the U-shaped curve in RT and accuracy was independent of preparation initiation time, consistent with the motivation-insensitive preparation hypothesis.

### 3.4.1 Motivational effects in task switching

There are two obvious forms of motivational effects in task switching: effort minimization and the speed-accuracy tradeoff. The impact on behavior of effort minimization strategies has previously been considered in a wide range of literatures (e.g., Ballard et al., 1997; Bettman et al., 1990; Botvinick et al., 2009; Droll & Hayhoe, 2007; Johnson & Payne, 1985; Kool et al., 2010; McGuire & Botvinick, 2010; Shah & Oppenheimer, 2008; Wilcox, 1993, 1994). In the task switching literature, De Jong (2000) and Nieuwenhuis and Monsell (2002) posited that mental effort associated with preparing for a task switch drives residual switch costs. Here I have argued against that idea.

However, there is yet another potential source of effort minimization effects in task switching, posited by Yeung and Monsell (2003) and later Savine and Braver (2010). These authors argued for an effort-performance tradeoff on the strength of task preparation. That is, they argued that task preparation is graded, with more preparation requiring more effort but translating to higher response process drift rate — that is, to faster and more accurate responses. Yeung and Monsell (2003) argued that this effort-drift tradeoff explains the asymmetry in switch costs when switching between tasks of unequal familiarity. Savine and Braver (2010) demonstrated this effort-drift tradeoff more directly in showing that a categorical motivational manipulation during task switching can increase both speed and accuracy. There is also a suggestion from those data that residual switch costs are reduced by this manipulation, although the operationalization of residual switch costs was a blockwise contrast between all trials in task switching blocks and all trials in single-task blocks, confounding precise inter-

pretation of the results. It is nonetheless surprising, considering the present results, that Savine and Braver elicited increased speed and accuracy, given that they also used a deadline paradigm that was similar to that employed here. It remains for future work to explore which types of motivational manipulations elicit a standard speed-accuracy tradeoff, as I observed in the present experimental data, and which types of motivational manipulations elicit this surprising increase in both speed and accuracy.

### 3.4.2 Model Assumptions & Alternatives

Although the model's qualitative pattern of behavior is robust to particular model forms, the conclusions that I drew following quantitative model selection with the Experiment 1 data depend on the form of the models. That is, minor differences in modeling choices could, in principle, lead to different results. This problem follows necessarily from precise model specifications, and is one reason that Experiment 2, which provided model-free support for the motivation-insensitive preparation hypothesis, was important. Nonetheless, it is is desirable, but beyond the scope of the present work, to consider the impact of alternative assumptions on the quantitative model selection results. I describe some potential alternatives below.

I have assumed that the preparatory process duration is an exponentially distributed r.v. This leads the model to reproduce the mixture properties of residual switch costs (e.g., De Jong, 2000; Nieuwenhuis & Monsell, 2002). Recall that this means that long-CSI switch RT distributions are mixtures of short-CSI switch and long-CSI repeat RT distributions. An alternative parametric form that would exactly reproduce this property is one in which preparatory process duration is constant, and preparation begins, randomly on each trial, either at the beginning of the CSI or at the time of target stimulus onset. In fact, this alternative model would be very similar to that employed here, with the major difference being that the shape of the switch

trial RT distribution would be somewhat different. Beyond these two models, there are many potential forms for the preparatory process duration, although most do not exactly reproduce the mixture phenomenon. Note that the mixture phenomenon, while an influential idea in the literature, has yet to be examined at the level of detail that would discriminate between these models, and indeed, at least one paper has argued that the long-CSI switch RT distribution is not the proposed mixture (Brown, Lehmann, & Poboka, 2006).

### 3.4.3 Conclusion

The residual switch cost is frequently studied as an experimental window into the apparent limits of human multitasking ability. This phenomenon has classically been interpreted as reflecting rigid limitations due to the structure of cognition. An alternative idea, that this limitation may be amenable to analysis and manipulation in terms of motivation and mental effort, has produced promising yet inconclusive results. Here I have presented a formal rational analysis (Anderson, 1990, 1991) of the task switching paradigm, and have concluded that, contrary to previous interpretations, model-based analysis shows that there is little evidence that motivation drives residual switch costs. Although this may be somewhat disappointing in that it eliminates a method of manipulating task preparation effects, the proposed framework affords a highly customizable platform for specifying and testing future hypotheses about motivational effects in the task switching paradigm.

## 3.5 Supporting Mathematical Details

### 3.5.1 Rational Analysis of Basic Task Switching

In general, I assume that participants maximize rewards (i.e., negative costs) under uncertainty and other constraints. To maximize generality of these results, I use the

expected *discounted* future reward optimality criterion, meaning that participants maximize the expected sum over future rewards, discounted by the delays before rewards are experienced. Behavioral predictions then follow from the RT and accuracy joint distribution implied by optimal decisions. The analyses presented above use undiscounted future reward optimality criterion, which is a special case of the discounted future reward criterion results presented here. For clarity, the results for the undiscounted special case are presented in the final section of this appendix.

In this section, I analyze switch and repeat trials in a basic task switching paradigm. Analysis of particular modifications of the paradigm, including the deadline paradigm described in the text above, are given as special cases in subsequent sections.

### Switch and Repeat Trial Rewards

As described in the main text, there are different expressions for the reward in switch and repeat trials, discounted from the time that the cue is first perceived. Conditional on RT and accuracy, switch trial reward is given by:

$$\begin{aligned} R_s|T,A &= c_l e^{-\alpha\tau} + \int_\tau^I c_m e^{-\alpha u} du + e^{-\alpha I}\left(\int_0^T (c_m + c_t)e^{-\alpha u} du + (1-A)c_e e^{-\alpha T}\right) \\ &= (c_l + \frac{c_m}{\alpha})e^{-\alpha\tau} + e^{-\alpha I}\left(\frac{c_t}{\alpha} - \frac{c_m + c_t}{\alpha}e^{-\alpha T} + (1-A)c_e e^{-\alpha T}\right) \end{aligned} \tag{3.1}$$

where $T$ is switch trial RT, $A$ indicates response accuracy, $\tau$ is preparation initiation time, $I$ is the duration of the CSI, $\alpha$ is the discount rate $(1/s)$, $c_l$ is the cost of initiating task preparation (\$), $c_m$ is the cost of maintaining task preparation $(\$/s)$, $c_t$ is the time or opportunity cost associated with participating in the experiment and with satisfying the desire to respond as quickly as possible $(\$/s)$, and $c_e$ is the cost associated with the desire to respond as accurately as possible (\$). To compute the expectation, I can condition on $T$ and $A$, which are independent from each other, as

described below. The expected switch trial reward is:

$$\mathbb{E}[R_s] = \left(c_l + \frac{c_m}{\alpha}\right) e^{-\alpha\tau} + e^{-\alpha I}\left(Q_1 \mathbb{E}\left[e^{-\alpha T}\right] + \frac{c_t}{\alpha}\right) \tag{3.2}$$
$$Q_1 = c_e \mathbb{P}\{A = 1\} - \frac{c_m + c_t}{\alpha}$$

Repeat trial reward depends on which repeat trial strategy participants use. They may choose to maintain the appropriate, already prepared task throughout the CSI, or they may choose instead to forget the appropriate task and then prepare it *de novo* at some point during the CSI. The latter choice makes repeat trials identical to switch trials. I denote this choice by the variable $\pi$, which takes on values 1 (maintain) or 3 (forget and re-prepare). Each of these strategies leads to a different expression for the repeat trial reward:

$$\begin{aligned} R_{r,1}|U, A &= \int_0^I c_m e^{-\alpha u} du + e^{-\alpha I}\left(\int_0^U (c_m + c_t) e^{-\alpha u} du + (1 - A) c_e e^{-\alpha U}\right) \\ &= \frac{c_m}{\alpha} + e^{-\alpha I}\left(\frac{c_t}{\alpha} - \frac{c_m + c_t}{\alpha} e^{-\alpha U} + (1 - A) c_e e^{-\alpha U}\right) \\ R_{r,3} &= R_s \end{aligned}$$

where $U$ is the basic DDM response duration. The expectations of each expression are:

$$\begin{aligned} \mathbb{E}[R_{r,1}] &= \frac{c_m}{\alpha} + e^{-\alpha I}\left(Q_1 \mathbb{E}\left[e^{-\alpha U}\right] + \frac{c_t}{\alpha}\right) \\ \mathbb{E}[R_{r,3}] &= \mathbb{E}[R_s] \end{aligned}$$

where $Q_1$ is defined as above, and $\mathbb{E}[R_s]$ is given in (3.2). Finally, the pieces can be

combined into an expression for the expected repeat trial reward:

$$\mathbb{E}[R_r] = \sum_{i \in \{1,3\}} \mathbb{I}\{\pi = i\}\mathbb{E}[R_{r,i}] \tag{3.3}$$

The expected switch and repeat trial reward functions depend on the participant's choice of preparation initiation time, $\tau$, repeat trial strategy, $\pi$, and implicitly on DDM threshold, $z$. Therefore, in the motivation-sensitive preparation model, participants solve the three-way optimization problem,

$$(\tau^*, z^*, \pi^*) \in \arg\max_{\tau,z,\pi} \{\mathbb{E}[R_s] + \mathbb{E}[R_r]\}. \tag{3.4}$$

In the motivation-insensitive preparation model, participants solve only for $(z^*, \pi^*)$, with $\tau$ being a free parameter of the model rather than a controlled quantity. Note that $\pi$ is discrete and takes only two values, so optimizing over this is easy. In contrast, $z$ is continuous and there is no closed form solution for its optimum value — it can be optimized numerically. Given $\pi$ and $z$, the optimization over continuous $\tau$ has a closed form solution, derived below. Note that expected rewards depend on response distributions ($T$, $U$, and $A$), which also depend on the optimized variables. Therefore, before solving for the optimal preparation initiation time, I describe some of the properties of the basic response processes.

### Basic DDM Equations & Properties

The DDM is a model of two-alternative-forced-choice decisions based on a form of one-dimensional Brownian motion. The moving particle maps to the process of accumulating evidence for one alternative or the other. Evidence accumulates over time as a drift signal, $\mu t$, plus zero-centered Gaussian noise with variance $\sigma^2 t$. The moving particle is bounded by absorbing barriers, known as decision thresholds, on either side of its initial position. The standard assumption is that participants choose the deci-

sion threshold so as to maximize some reward function. This physical model serves as a joint distribution over RT and accuracy. RT is modeled as the "first passage time" at which the particle first reaches one of the thresholds plus a constant non-decision time. Accuracy is modeled by which threshold the particle hits.

I use a relatively simple version of the DDM in which the particle's initial position is zero, with symmetric decision thresholds at $+/-z$ (chosen by participants). Moreover, any change in the noise parameter can be absorbed by rescaling the remaining parameters, so noise is fixed at 0.1 without loss of generality, although the letter $\sigma$ remains in the equations for clarity.

Finally, in the task switching application it is necessary to further break down the non-decision time component of RT into two separate parts. $T_1$ denotes pre-decision perceptual processes (e.g., retinal and early visual processes), and $T_0$ to denote post-decision motor processes (e.g., moving a finger toward the response button). This decomposition is necessary, because I assume that early perceptual processes occur irrespective of task preparedness. In contrast, as noted in the main text evidence accumulation is assumed to begin only after task preparation completes. Therefore, non-decision time must be decomposed into one part that is unaffected by preparation, and another that is. This is made more precise in (3.10), below.

$U$ denotes the first passage time plus the non-decision components, which is RT in the basic DDM model. Bernoulli variable $A$ denotes accuracy. $A$ is 1 if the particle passes the positive (correct) threshold, or zero if the particle passes the negative (incorrect) threshold. Note that both of these random variables implicitly depend on the chosen decision threshold. The first useful fact about the model is that RT and accuracy are independent given model parameters and threshold. The joint distribution factorizes as $\mathbb{P}\{U \in dt, A = a\} = \mathbb{P}\{U \in dt\}\mathbb{P}\{A = a\}$. Therefore it is sufficient to compute the marginal probabilities. Beginning with RT, the probability

density function (p.d.f.) is (Ratcliff, 1978, Eq. A9; see final section, below):

$$\mathbb{P}\{U \in dt\} = \begin{cases} C(\mu, z) \sum_{n=0}^{\infty} C_n \exp\{-A_n (t - (T_0 + T_1))\}, & T_0 + T_1 \leq t \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

$$C(\mu, z) = \frac{\pi\sigma^2}{4z^2} 2\cosh(\frac{\mu z}{\sigma^2})$$

$$C_n = (-1)^n(2n+1)$$

$$A_n = (2n+1)^2 \frac{(\pi\sigma)^2}{8z^2} + \frac{\mu^2}{2\sigma^2}$$

The probability distribution for accuracy is:

$$\mathbb{P}\{A = 1\} = \frac{1 - e^B}{e^{-B} - e^B}, \quad \mathbb{P}\{A = 0\} = \frac{e^{-B} - 1}{e^{-B} - e^B}, \quad B = \frac{2\mu z}{\sigma^2} \tag{3.6}$$

Present analyses require a few more properties of the basic DDM model, including the expected RT,

$$\mathbb{E}[U] = \frac{z}{\mu} \tanh\left(\frac{\mu z}{\sigma^2}\right) + T_0 + T_1, \tag{3.7}$$

and the Laplace transform of the p.d.f., which follows from the result provided by Smith (1990):

$$\begin{aligned} \mathbb{E}\left[e^{-\alpha U}\right] &= \int_0^{\infty} e^{-\alpha t} \mathbb{P}\{U \in dt\} \\ &= e^{-\alpha(T_0+T_1)} 2\cosh\left(\frac{\mu z}{\sigma^2}\right) \frac{\sinh(Dz)}{\sinh(2Dz)} \\ D &= \sqrt{\frac{2\alpha}{\sigma^2} + \frac{\mu^2}{\sigma^4}} \end{aligned} \tag{3.8}$$

Finally, the incomplete Laplace transform of the p.d.f. will be required:

$$\begin{aligned}
\psi_U(\alpha, x) &= \int_0^x e^{-\alpha t}\mathbb{P}\{U \in dt\} \\
&= \int_{T_0+T_1}^x e^{-\alpha t} C(\mu, z) \sum_{n=0}^{\infty} C_n \exp\{-A_n(t-(T_0+T_1))\}dt \\
&= C(\mu, z) \sum_{n=0}^{\infty} \int_{T_0+T_1}^x C_n \exp\{-(\alpha+A_n)(t-(T_0+T_1))\}dt \\
&= e^{-\alpha(T_0+T_1)} C(\mu, z) \sum_{n=0}^{\infty} \frac{-C_n}{\alpha+A_n} \left[\exp\{-(\alpha+A_n)(x-(T_0+T_1))\} - 1\right] \\
&= \mathbb{E}\left[e^{-\alpha U}\right] - e^{-\alpha(T_0+T_1)} C(\mu, z) \sum_{n=0}^{\infty} \frac{C_n}{\alpha+A_n} \exp\{-(\alpha+A_n)(x-(T_0+T_1))\}
\end{aligned} \tag{3.9}$$

This section presented some facts about the DDM RT and accuracy distributions, which depend on the participant's choice of decision threshold, $z$. The next section describes how the basic DDM response model interacts with the participant's choice of preparation initiation time on switch trials.

### Combining DDM with Preparation

Since I assume that the basic DDM response process does not begin until the preparatory process completes, response accuracy is not directly affected by preparation. Therefore the probability of an accurate response is the same here as with the basic DDM.

With regard to RT, the basic principle is that the switch trial RT distribution is a mixture of early and late preparation trials. In early preparation trials, the task is fully prepared before target onset and early perception, and the DDM begins at the normal time, leading to RTs that are indistinguishable from repeat trial RTs. In late preparation trials, the task is not prepared prior to target onset and early perception, leading to an RT that is the sum of the DDM *and* the remaining preparatory process

duration. Moreover, due to the memoryless property of the exponential distribution, the remaining preparatory process duration is independent and identically distributed to the original preparatory process duration.

$T$ denotes switch trial RT. Conditional on preparatory process duration and basic DDM response time, $T$ is expressed deterministically:

$$T|S,U,V = \mathbb{I}\{S \leq I + T_1 - \tau\}U + \mathbb{I}\{S > I + T_1 - \tau\}\,(U + V) \tag{3.10}$$

where $S$ and $V$ are independent and identically distributed (i.i.d.) random variables drawn from the exponential preparatory process duration distribution, and $\mathbb{I}\{\cdot\}$ is one if its argument is true, zero otherwise. On the right-hand side, the left term expresses the early preparation case, and the right term expresses the late preparation case. In the late preparation case, the memoryless preparatory process "starts over" with the new exponential random variable, $V$. From this definition, the switch RT density

function can be derived by conditioning on when preparation completes:

$$\mathbb{P}\{T \in dt\} = \begin{cases} \mathbb{P}\{S \leq I + T_1 - \tau\}\mathbb{P}\{U \in dt\} \\ \quad + \mathbb{P}\{S > I + T_1 - \tau\}\mathbb{P}\{U + V \in dt\}, & T_0 + T_1 < t \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{P}\{S \leq I + T_1 - \tau\} = \begin{cases} 1 - e^{\lambda(I+T_1-\tau)}, & \tau < I + T_1 \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

$$\mathbb{P}\{U + V \in dt\} = \begin{cases} \lambda C(\mu, z) \sum_{n=0}^{\infty} \frac{C_n}{\lambda - A_n} \\ \quad \cdot \Big( \exp\{-A_n(t - (T_0 + T_1))\} \\ \quad - \exp\{-\lambda(t - (T_0 + T_1))\} \Big), & (T_0 + T_1) \leq t \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda$ is the rate parameter for the exponential preparation process duration ($1/s$). There is no closed form solution for the p.d.f. without an infinite series, as indeed there is no closed form solution for the p.d.f. of the basic DDM response duration given in (3.5). Luckily, the analyses that follow will rely heavily on two functionals of this p.d.f. for which I do have closed form solutions, namely the expectation, $\mathbb{E}[T]$, and Laplace transform, $\mathbb{E}[e^{-\alpha T}]$. I can obtain these by inspection of the definition of $T$ given in (3.10). Noting that, for any $X$ and $Y$, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, I have:

$$\begin{aligned} \mathbb{E}[T] &= \mathbb{P}\{S \leq I + T_1 - \tau\}\mathbb{E}[U] + \mathbb{P}\{S > I + T_1 - \tau\}\mathbb{E}[U + V] \\ &= \mathbb{P}\{S \leq I + T_1 - \tau\}\mathbb{E}[U] + \mathbb{P}\{S > I + T_1 - \tau\}\left(\mathbb{E}[U] + \mathbb{E}[V]\right) \\ \mathbb{E}[V] &= \frac{1}{\lambda} \end{aligned} \tag{3.12}$$

The expected hitting time of the basic DDM is given in (3.7). Finally, noting that, for independent r.v.s $X$ and $Y$, $\mathbb{E}[e^{-\alpha(X+Y)}] = \mathbb{E}[e^{-\alpha X}]\mathbb{E}[e^{-\alpha Y}]$, the Laplace transform of the switch trial RT is given by:

$$\mathbb{E}\left[e^{-\alpha T}\right] = \mathbb{P}\{S \leq I + T_1 - \tau\}\mathbb{E}\left[e^{-\alpha U}\right] + \mathbb{P}\{S > I + T_1 - \tau\}\mathbb{E}\left[e^{-\alpha U}\right]\mathbb{E}\left[e^{-\alpha V}\right] \tag{3.13}$$

$$= \left(\mathbb{P}\{S \leq I + T_1 - \tau\} + \mathbb{P}\{S > I + T_1 - \tau\}\mathbb{E}\left[e^{-\alpha V}\right]\right)\mathbb{E}\left[e^{-\alpha U}\right] \tag{3.14}$$

$$+ \mathbb{P}\{S > I + T_1 - \tau\}\mathbb{E}\left[e^{-\alpha V}\right]$$

$$\mathbb{E}\left[e^{-\alpha V}\right] = \frac{\lambda}{\alpha + \lambda}$$

The incomplete Laplace transform of the switch trial RT density will also be necessary:

$$\psi_T(\alpha, x) = \int_0^x e^{-\alpha t}\mathbb{P}\{T \in dt\}$$

$$= \mathbb{P}\{S \leq I + T_1 - \tau\}\psi_U(\alpha, x) + \mathbb{P}\{S > I + T_1 - \tau\}\psi_{U+V}(\alpha, x) \tag{3.15}$$

$$\psi_{U+V}(\alpha, x) = e^{-\alpha(T_0+T_1)}\lambda C(\mu, z)\sum_{n=0}^{\infty}\frac{C_n}{A_n - \lambda}F_n$$

$$F_n = \frac{\exp\{-(A_n + \alpha)(x - (T_0 + T_1))\} - 1}{A_n + \alpha} \tag{3.16}$$

$$- \frac{\exp\{-(\alpha + \lambda)(x - (T_0 + T_1))\} - 1}{\alpha + \lambda}$$

In this section, the consequences of combining the basic DDM response process with assumptions about switch trial preparation were considered. I can now return to the problem of finding the optimal preparation initiation time.

### Optimal Preparation Initiation Time

Optimal preparation initiation time can now be derived as a function of DDM threshold. Note that optimal preparation initiation time does not depend on repeat trial strategy (the third optimized variable). Rather, in the case that participants choose

to forget and re-prepare on repeat trials, this derivation holds for repeat as well as switch trials. To solve for the optimal preparation initiation time, take the partial derivative of the expected switch trial reward given in (3.2):

$$\begin{aligned}\frac{\partial}{\partial\tau}\mathbb{E}[R_s] &= -(\alpha c_l + c_m)e^{-\alpha\tau} + e^{-\alpha I}\left(Q_1\frac{\partial}{\partial\tau}\mathbb{E}\left[e^{-\alpha T}\right]\right)\\ \frac{\partial}{\partial\tau}\mathbb{E}\left[e^{-\alpha T}\right] &= -\frac{\alpha\lambda}{\alpha+\lambda}e^{-\lambda(I+T_1)}\mathbb{E}\left[e^{-\alpha U}\right]e^{\lambda\tau}.\end{aligned} \tag{3.17}$$

The partial derivative is simply a sum of two exponentials, and so to find the critical point, set it to zero, rearrange and take the log of both sides, yielding an analytic expression for the optimal preparation initiation time:

$$\tau^*(z) = I + \frac{\lambda}{\alpha+\lambda}T_1 + \frac{1}{\alpha+\lambda}\log\left[\frac{\alpha c_l + c_m}{-Q_1\frac{\alpha\lambda}{\alpha+\lambda}\mathbb{E}\left[e^{-\alpha U}\right]}\right] \tag{3.18}$$

Thus, for a given $z$, there is a unique $\tau^*(z)$.

Recall that optimizing over repeat trial strategy is a simple discrete problem, whereas there is no analytic solution for $z^*$ however, this variable can be numerically optimized when working with the model. After this one-dimensional numerical optimization problem is solved (with the nested two-dimensional analytic optimization problem), the optimization problem set forth in equation (3.4) is solved.

### 3.5.2 Special Cases

With the caveat that actual preparation initiation can never be negative, the basic analysis described above is sufficient to model residual switch costs (e.g., Rogers & Monsell, 1995), and the failure to engage property (e.g., De Jong, 2000; Nieuwenhuis & Monsell, 2002). Analyses for the special cases of other experiments considered in the main text are presented below.

### Probabilistic Switch Cues

Dreisbach and Haider (2006) used probabilistic switch cues, allowing them to parametrically manipulate from trial to trial the probability of repeating the previous task versus switching to the alternate task. To analyze this, first note that, at the moment that the cue is perceived, switch and repeat trials are equivalent, since the actual switch/repeat difference will not be revealed until later, at target onset. Switch and repeat trial rewards can then be combined into a single function that is parameterized by the probability of repeating, $\rho$. The previous expected switch and repeat trial reward expressions can be seen as special cases of this function for which $\rho = 0$ and $\rho = 1$, respectively.

Next expand the role of the discrete "repeat trial strategy" variable, $\pi$, because choosing a strategy based on the probability of repeating now applies to all trials. Participants may now reasonably choose from four strategies: (1) maintain, (2) forget and prepare for switch, (3) forget and re-prepare for repeat, and (4) forget and prepare only when the actual task is revealed at target onset. Define the combined reward function as a sum of auxiliary functions indexed by strategy:

$$\mathbb{E}\left[R\right] = \sum_{i=1}^{4} \mathbb{I}\{\pi = i\}\mathbb{E}[R_i]$$

$$\mathbb{E}\left[R_1\right] = -\frac{c_m}{\alpha}\left(e^{-\alpha I} - 1\right) + \rho e^{-\alpha I}\left(Q_1\mathbb{E}\left[e^{-\alpha U}\right] + \frac{c_m + c_t}{\alpha}\right) + (1-\rho)\mathbb{E}[R_4]$$

$$\mathbb{E}\left[R_2\right] = c_l e^{-\alpha\tau} - \frac{c_m}{\alpha}\left(e^{-\alpha I} - e^{-\alpha\tau}\right) + (1-\rho)e^{-\alpha I}\left(Q_1\mathbb{E}\left[e^{-\alpha T}\right] + \frac{c_m + c_t}{\alpha}\right) + \rho\mathbb{E}[R_4]$$

$$\mathbb{E}\left[R_3\right] = c_l e^{-\alpha\tau} - \frac{c_m}{\alpha}\left(e^{-\alpha I} - e^{-\alpha\tau}\right) + \rho e^{-\alpha I}\left(Q_1\mathbb{E}\left[e^{-\alpha T}\right] + \frac{c_m + c_t}{\alpha}\right) + (1-\rho)\mathbb{E}[R_4]$$

$$\mathbb{E}\left[R_4\right] = e^{-\alpha I}\left(c_l + Q_1\mathbb{E}\left[e^{-\alpha T(I,z)}\right] + \frac{c_m + c_t}{\alpha}\right)$$

Given $\pi = 1$ or $\pi = 4$, $\tau$ has no effect on the expected reward. Given $\pi = 2$ or $\pi = 3$,

a slight generalization of my basic result from (3.18) produces:

$$\tau^*_{\pi=2}(z) = I + \frac{\lambda}{\alpha+\lambda}T_1 + \frac{1}{\alpha+\lambda}\log\left[\frac{\alpha c_l + c_m}{Q_1(1-\rho)\frac{-\alpha\lambda}{\alpha+\lambda}\mathbb{E}\left[e^{-\alpha U}\right]}\right],$$
$$\tau^*_{\pi=3}(z) = I + \frac{\lambda}{\alpha+\lambda}T_1 + \frac{1}{\alpha+\lambda}\log\left[\frac{\alpha c_l + c_m}{Q_1\rho\frac{-\alpha\lambda}{\alpha+\lambda}\mathbb{E}\left[e^{-\alpha U}\right]}\right].$$

As above, the optimization over $\pi$ is discrete and tractable, whereas the optimization over $z$ is intractable and must be approximated numerically.

### Special Case: Deadline Paradigm

The deadline paradigm requires adding a term to the switch trial reward function given in (3.2):

$$\begin{aligned} R_s|T, A &= c_l e^{-\alpha\tau} + \int_\tau^I c_m e^{-\alpha u}du + e^{-\alpha I}\Bigg(\int_0^T (c_m + c_t)e^{-\alpha u}du \\ &\qquad + (1-A)(c_e + c_p)e^{-\alpha T} + \mathbb{I}\{T \le D, A = 1\}r_t e^{-\alpha T}\Bigg) \\ &= (c_l + \frac{c_m}{\alpha})e^{-\alpha\tau} + e^{-\alpha I}\Bigg(\frac{c_t}{\alpha} - \frac{c_m + c_t}{\alpha}e^{-\alpha T} + (1-A)(c_e + c_p)e^{-\alpha T} \\ &\qquad + \mathbb{I}\{T \le D\}Ar_t e^{-\alpha T}\Bigg), \end{aligned}$$

where $D$ denotes the response deadline ($s$), $c_p$ denotes the monetary error penalty ($), and $r_t$ denotes the monetary reward for accurate responses made before the deadline ($). The expected switch trial reward becomes:

$$\begin{aligned} \mathbb{E}[R_s] &= \left(c_l + \frac{c_m}{\alpha}\right)e^{-\alpha\tau} + e^{-\alpha I}\left(Q_1\mathbb{E}\left[e^{-\alpha T}\right] + Q_2\psi_T(\alpha, D) + \frac{c_t}{\alpha}\right) \qquad (3.19) \\ Q_1 &= (c_e + c_p)\mathbb{P}\{A = 0\} - \frac{c_m + c_t}{\alpha} \\ Q_2 &= r_t\mathbb{P}\{A = 1\} \end{aligned}$$

To compute the optimal preparation initiation time, take the partial derivative of the expected switch trial reward once again,

$$\frac{\partial}{\partial\tau}\mathbb{E}[R_s] = -(\alpha c_l + c_m)e^{-\alpha\tau} + e^{-\alpha I}\left(Q_1\frac{\partial}{\partial\tau}\mathbb{E}\left[e^{-\alpha T}\right] + Q_2\frac{\partial}{\partial\tau}\psi_T(\alpha, D)\right), \quad (3.20)$$

where $\frac{\partial}{\partial\tau}\mathbb{E}\left[e^{-\alpha T}\right]$ is given in (3.17). The incomplete Laplace transform, $\psi_T(\alpha, D)$, was given in a simple form in (3.15). However a slightly different derivation here provides more efficient numerical computation of the partial derivative:

$$\begin{aligned}
\frac{\partial}{\partial\tau}\psi_T(\alpha, D) &= \int_0^D \frac{\partial}{\partial\tau}\mathbb{P}\{T \in dt\} \\
&= \int_{T_0+T_1}^D \frac{\partial}{\partial\tau}\mathbb{P}\{T \in dt\} \\
&= \left(\lambda e^{-\lambda(I+T_1)}\int_{T_0+T_1}^D \phi(t; z)dt\right)e^{\lambda\tau} \quad (3.21) \\
\phi(t; z) &= -C(\mu, z)\sum_{n=0}^{\infty} C_n e^{-A_n(t-(T_0+T_1))} \quad (3.22) \\
&\quad + \lambda C(\mu, z)\sum_{n=0}^{\infty}\frac{C_n}{\lambda - A_n}\left[e^{-A_n(t-(T_0+T_1))} - e^{-\lambda(t-(T_0+T_1))}\right].
\end{aligned}$$

As above, I can solve for the critical point, $\tau^*$, as a function of $z$:

$$\tau^*(z) = I + \frac{\lambda}{\alpha+\lambda}T_1 + \frac{1}{\alpha+\lambda}\log\left[\frac{\alpha c_l + c_m}{\lambda\left(Q_1\frac{-\alpha}{\alpha+\lambda}\mathbb{E}\left[e^{-\alpha U}\right] + Q_2\int_{T_0+T_1}^D e^{-\alpha t}\phi(t; z)dt\right)}\right]$$

As with the switch trial reward, the deadline paradigm introduces another term

into the repeat trial reward:

$$
\begin{aligned}
R_{r,1}|U,A &= \int_0^I c_m e^{-\alpha u} du + e^{-\alpha I}\Bigg( \int_0^U (c_m + c_t) e^{-\alpha u} du \\
&\qquad + (1-A)(c_e + c_p)e^{-\alpha U} + \mathbb{I}\{U \leq D, A = 1\} r_t e^{-\alpha U} \Bigg) \\
&= \frac{c_m}{\alpha} + e^{-\alpha I}\Bigg( \frac{c_t}{\alpha} - \frac{c_m + c_t}{\alpha} e^{-\alpha U} + (1-A)(c_e + c_p)e^{-\alpha U} \\
&\qquad + \mathbb{I}\{U \leq D\} A r_t e^{-\alpha U} \Bigg) \\
R_{r,3} &= R_s
\end{aligned}
\tag{3.23}
$$

The expectations can again be computed for each piece, and combined into an expression for the repeat trial expected reward function:

$$
\begin{aligned}
\mathbb{E}[R_r] &= \sum_{\pi \in \{1,3\}} \mathbb{I}\{\pi = i\} \mathbb{E}[R_{r,1}] \\
\mathbb{E}[R_{r,1}] &= \frac{c_m}{\alpha} + e^{-\alpha I}\left( Q_1 \mathbb{E}\left[e^{-\alpha U}\right] + Q_2 \psi_U(\alpha, D) + \frac{c_t}{\alpha} \right) \\
\mathbb{E}[R_{r,3}] &= \mathbb{E}[R_s],
\end{aligned}
$$

### 3.5.3 Alternate Optimality Criterion: Undiscounted Expected Future Reward

The analyses above were conducted using the expected, discounted, future reward criterion. In addition to facilitating analyses that lie outside the considerations of this paper (as mentioned above in the Discussion section), the inclusion of a discount factor takes the first step toward relating this modeling framework to literatures on temporal discounting. It also contains the simpler undiscounted expected future reward criterion analysis as the special case when $\alpha = 0$. For completeness, analyses of this special case are presented here, for the deadline paradigm. Note that the case

of a basic task switching experiment can be recovered from this by setting $D$ and $c_p$ to zero.

As with the discounted case, I assume that participants solve the optimization problem set out in equation (3.4). The expected reward functions are now different however, reflecting the undiscounted optimality criterion. Begin with the switch trial reward:

$$R_s|T, A = c_\ell + c_m(I - \tau) + (c_m + c_t)T + (1 - A)(c_e + c_p) + A\mathbb{I}\{T \leq D\}r_t.$$

The expected switch trial reward is then:

$$\mathbb{E}[R_s] = c_\ell + c_m(I - \tau) + (c_m + c_t)\mathbb{E}[T] + \mathbb{P}\{A = 0\}(c_e + c_p) + \mathbb{P}\{A = 1\}\mathbb{P}\{T \leq D\}r_t.$$

Note that $\mathbb{E}[T]$ is given in (3.12) above, whereas $\mathbb{P}\{T \leq D\}$ may be obtained by similar logic. As with the discounted case, can solve analytically for optimal preparation initiation time as a function of DDM decision threshold. The following partial derivatives are necessary:

$$\begin{aligned}
\frac{\partial}{\partial\tau}\mathbb{E}[R_s] &= -c_m + (c_m + c_t)\frac{\partial}{\partial\tau}\mathbb{E}[T] + \mathbb{P}\{A = 1\}\frac{\partial}{\partial\tau}\mathbb{P}\{T \leq D\}r_t \\
\frac{\partial}{\partial\tau}\mathbb{E}[T] &= e^{-\lambda(I+T_1-\tau)} \\
\frac{\partial}{\partial\tau}\mathbb{P}\{T \leq D\} &= \lambda e^{-\lambda(I+T_1-\tau)}\int_{T_0+T_1}^{D}\phi(u)du \\
\int_{T_0+T_1}^{D}\phi(u)du &= \left[\int_{T_0+T_1}^{D}\mathbb{P}\{U + V \in du\} - \int_{T_0+T_1}^{D}\mathbb{P}\{U \in du\}\right]
\end{aligned}$$

This leads to the following optimal preparation initiation time as a function of DDM decision threshold:

$$\tau^*(z) = I + T_1 + \frac{1}{\lambda}\log\left[\frac{c_m}{c_m + c_t + \lambda\mathbb{P}\{A = 1\}r_t\int_{T_0+T_1}^{D}\phi(u)du}\right] \tag{3.24}$$

Repeat trial reward is again indexed by repeat trial strategy, $\pi$:

$$R_{r,1}|U, A = c_m I + (c_m + c_t)U + (1 - A)(c_e + c_p) + A\mathbb{I}\{U \leq D\}r_t$$
$$R_{r,3}|T, A = R_s|T, A.$$

with expectations:

$$\mathbb{E}[R_{r,1}] = c_m I + (c_m + c_t)\mathbb{E}[U] + (1 - A)(c_e + c_p) + A\mathbb{P}\{U \leq D\}r_t$$
$$\mathbb{E}[R_{r,3}] = \mathbb{E}[R_s].$$

Expected repeat trial reward is:

$$\mathbb{E}[R_r] = \sum_{i \in \{1,3\}} \mathbb{I}\{\pi = i\}\mathbb{E}[R_{r,i}].$$

### 3.5.4 Obtaining My DDM Equations From Ratcliff (1978)

To derive the basic DDM RT p.d.f. given in Eq. (3.5), begin with Ratcliff's (1978) Eq. A9:

$$\mathbb{P}\{U \in dt, X(t) = 0\} = \frac{\pi s^2}{a^2} e^{-(z\xi/s^2)} \sum_{k=1}^{\infty} k \sin\left(\frac{\pi z k}{a}\right) \exp\left\{-\frac{1}{2}\left(\frac{\xi^2}{s^2} + \frac{\pi^2 k^2 s^2}{a^2}\right) t\right\}. \tag{3.25}$$

Substitute $\mu = \xi$, $\sigma = s$ to accommodate my notation:

$$\mathbb{P}\{U \in dt, X(t) = 0\} = \frac{\pi \sigma^2}{a^2} e^{-(z\mu/\sigma^2)} \sum_{k=1}^{\infty} k \sin\left(\frac{\pi z k}{a}\right) \exp\left\{-\frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \frac{\pi^2 k^2 \sigma^2}{a^2}\right) t\right\}. \tag{3.26}$$

Replace $a$ with $2z$, because I assume barriers are symmetric about the initial position:

$$\mathbb{P}\{U \in dt, X(t) = 0\} = \frac{\pi\sigma^2}{4z^2} e^{-(z\mu/\sigma^2)} \sum_{k=1}^{\infty} k \sin\left(\frac{\pi k}{2}\right) \exp\left\{-\left(\frac{\mu^2}{2\sigma^2} + \frac{\pi^2 k^2 \sigma^2}{8z^2}\right) t\right\}. \tag{3.27}$$

Next, apply the change of variable $n = k/2 - 1/2$:

$$\mathbb{P}\{U \in dt, X(t) = 0\} = \frac{\pi\sigma^2}{4z^2} e^{-(z\mu/\sigma^2)} \sum_{n=0}^{\infty} (2n+1) \sin\left(\frac{\pi(2n+1)}{2}\right) \exp\{-A_n t\} \tag{3.28}$$

$$= \frac{\pi\sigma^2}{4z^2} e^{-(z\mu/\sigma^2)} \sum_{n=0}^{\infty} C_n \exp\{-A_n t\}, \tag{3.29}$$

$$A_n = \frac{\mu^2}{2\sigma^2} + \frac{\pi^2 (2n+1)^2 \sigma^2}{8z^2}, \tag{3.30}$$

$$C_n = (-1)^n (2n+1). \tag{3.31}$$

Next, form $\mathbb{P}\{U \in dt, X(t) = a\}$, as in Ratcliff's appendix, and combine the joints to form the p.d.f.:

$$\mathbb{P}\{U \in dt\} = \mathbb{P}\{U \in dt, X(t) = 0\} + \mathbb{P}\{U \in dt, X(t) = a\} \tag{3.32}$$

$$= C(\mu, z) \sum_{n=0}^{\infty} C_n \exp\{-A_n t\}, \tag{3.33}$$

$$C(\mu, z) = \frac{\pi\sigma^2}{4z^2} 2 \cosh\left(\frac{z\mu}{\sigma^2}\right). \tag{3.34}$$

Finally, the residual RT component shifts the p.d.f. to the right by $T_0 + T_1$:

$$\mathbb{P}\{U \in dt\} = C(\mu, z) \sum_{n=0}^{\infty} C_n \exp\{-A_n (t - (T_0 + T_1))\}, \tag{3.35}$$

which is Eq. (3.5), as intended.

# Chapter 4

# Confounds in fMRI Mapping of Context Representations

## 4.1 Introduction

Mapping the neural substrates of context representations, the component of task representations that is theorized to reside in PFC, is essential for testing aspects of GAT and Gating models. This has proven difficult to achieve with functional magnetic resonance imaging (fMRI), despite success with representation mapping in other representational domains. New fMRI methods have led to recent reports of success in this domain. In this chapter, I critically examine these reports in terms of a methodological confound that may not be widely understood. I conclude that the recent reports are in fact better interpreted as maps of a confounding task difficulty variable, rather than as maps of context representations *per se*. In this chapter, I discuss one simple method of addressing these confounds with existing methods. In the next chapters, I propose and apply newer methods that address the confounds in a more complete way. The recent reports discussed below have utilized novel, instructed tasks, and have typically labeled the resulting representations as "rule"

representations. I use that term in this chapter. In terms of the GAT variables that form the broader framework of this dissertation, it should be understood that a rule representation corresponds to a context representation as well as a novel stimulus-response mapping, and that the two representations are tightly coupled (confounded) in the datasets discussed below.

Multivariate pattern analysis (MVPA) is a recent development in fMRI that allows data analysts to detect relationships between experiment conditions and complex voxel activity patterns. One major application of MVPA is cognitive mapping, in which one assesses where certain information is represented in the brain. Although this has been a fundamental application of fMRI since its inception, MVPA theoretically and perhaps empirically yields greater sensitivity than do classical, general linear model (GLM) methods. This has led to reports in many domains of new representation mapping results that may have been impossible with the GLM. In two seminal examples, both Haynes and Rees (2005) and Kamitani and Tong (2005) reported using MVPA to detect visual stimulus orientation representations in primary visual cortex (but see Freeman, Brouwer, Heeger, & Merriam, 2011). Later, Haynes et al. (2007) reported that intention to perform either addition or subtraction was represented in medial prefrontal cortex. Peelen, Atkinson, and Vuilleumier (2010) reported that perceived emotional states of other people were represented in medial prefrontal cortex and superior temporal sulcus, and Carlin, Calder, Kriegeskorte, Nili, and Rowe (2011) reported that perceived gaze direction of others (irrespective of viewing angle) was represented in superior temporal sulcus. Vickery, Chun, and Lee (2011) reported that the reinforcing or punishing value of outcomes was represented widely across gray matter voxels. Finally, as shall be discussed in more detail below, several authors have recently reported that task rules are represented in lateral prefrontal cortex and parietal areas (Bode & Haynes, 2009; Reverberi et al., 2011; Woolgar et al., 2011; Cole et al., 2011).

It is widely assumed that MVPA's increased mapping sensitivity arises from its ability to detect finely distributed representation patterns. Indeed, in regions where MVPA detects representations but the GLM does not, it is typical to conclude that not only do the identified brain regions represent the information in question, but that the representations are distributed at a fine spatial scale across voxels within the regions (e.g., Haxby et al., 2001; Kamitani & Tong, 2005; Reverberi et al., 2011). This conclusion follows from the fact that finely distributed patterns are very difficult for the GLM to detect in practice (e.g., Forman et al., 1995), but are detectable by MVPA in theory (Boynton, 2005; Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006; Kriegeskorte, Goebel, & Bandettini, 2006, 2007).[1] Unfortunately, I demonstrate in this chapter that across-subject counterbalancing has no effect in standard, classification-based MVPA mapping. That is, in classification-based MVPA mapping in general, all variables that are confounds at the single-subject level, such as order effects or individual differences in the response to experiment conditions, are reflected in the ultimate group results, irrespective of counterbalancing. This is in contrast to the GLM, in which across-subject counterbalancing has its intended effects, so that confounds and individual differences are eliminated during group tests, as expected. This admits the possibility that MVPA's increased mapping sensitivity is in fact spurious, and arises from confounds rather than detection of distributed representations.

I note that there are at least two major applications of MVPA: mapping, introduced above, and so-called mind reading, in which a scalar signal is estimated from wholebrain pattern activity and used to constrain behavioral models. Moreover, within mapping there are standard, classification-based variants and newer,

[1]The GLM can detect finely distributed patterns in theory. However, in practice imperfect inter-subject alignment and the burden of correcting for multiple comparisons across voxels lead to the practice of searching for spatially contiguous clusters of co-activated or co-deactivated (i.e., same-signed) voxels (Forman et al., 1995; Smith & Nichols, 2009). This dependence on spatially contiguous, same-signed activation effectively precludes the detection of fine patterns with the GLM, particularly in group tests.

representational similarity analysis (RSA: Kriegeskorte et al., 2008a) variants. The susceptibility to confounds that I describe here affects classification-based mapping in general, but affects other forms of MVPA to lesser degrees. Specifically, the conclusions of RSA are affected in some, but not all, cases. The conclusions of mind reading studies (e.g., Polyn, Natu, Cohen, & Norman, 2005; Johnson, McDuff, Rugg, & Norman, 2009) may be affected very little or perhaps not at all. I discuss implications for these other forms of MVPA in greater detail below.

I propose a method for controlling for these confounds, based on two passes of linear regression. I then present a case study, performing classification-based mapping of rule representations (i.e., flexibly updated stimulus-response associations). This case study is particularly relevant at this time, as there have been several recent articles reporting successful classification-based mapping of rule representations (Bode & Haynes, 2009; Cole et al., 2011; Woolgar et al., 2011; Reverberi et al., 2011), and it is possible that at least part of those results are due to confounds. I present novel experimental data in which I obtain rule representation maps similar to those recently reported, and obtain the same profile of success with MVPA and failure with the GLM. I then identify a confound between rule condition and task difficulty, as indexed by reaction time (RT). Critically, group behavioral ANOVA does not reveal this confound, because it tends to be of opposite sign in different subjects (i.e., it is a random individual difference). Moreover, I show that the practice of subtracting the across-voxel mean from classification examples (e.g., Cole et al., 2011) has virtually no effect on the results, and thus does not facilitate confound removal or interpretation. I then apply the proposed method of confound control based on linear regression, and show that this completely eliminates all rule mapping results. In contrast, it leaves a response direction representation map essentially unaffected, demonstrating specificity. This case study empirically demonstrates the general susceptibility of classification-based MVPA to confounds, and also suggests that recent rule mapping

results may be subject to reinterpretation in terms of a confound between rule and difficulty.

Finally, I note that, as is always the case with confounds, this issue does not demonstrate that classification-based MVPA mapping results are spurious. Rather, it reveals that experimental control is looser in MVPA mapping studies than is widely realized (or typically accepted in the field of experimental fMRI), and admits the possibility of alternative explanations in terms of less desirable alternative hypotheses. In other words, existing classification-based MVPA mapping results become more equivocal, but are not refuted.

## 4.2 Directional vs Directionless Single-Subject Summaries

In this section, I discuss in detail how classification-based MVPA mapping eliminates across-subject counterbalancing. I contrast this with the GLM, and discuss how other variants of MVPA are affected.

In GLM mapping analysis, separately for each subject and each voxel, a linear regression model specifying the experiment conditions is fit to the data. A contrast or linear combination of partial regression coefficients, specifying a hypothesized relationship between voxel activity in different conditions, then forms a summary statistic for that voxel and that subject. This score is directional in the sense that its deviations from zero reflect the voxel's activation or deactivation in response to the contrast: an above-zero score indicates that the voxel was activated, whereas a below-zero score indicates that the voxel was deactivated. These single-subject contrast scores can then be submitted to a group t-test in order to test the null hypothesis that, in the population of subjects, the particular voxel's activation is not significantly above (or below) zero.

In classification-based MVPA mapping analysis, separately for each subject and each region, a classifier algorithm is trained on part of the dataset to discriminate between experiment conditions, and is then tested on its ability to do so on a separate part of the data.[2] This yields a classification summary score for each subject and each region. This summary score is akin to a z-scored single-subject significance value, with more positive values indicating greater significance (i.e., a lower single-subject p-value). Critically, this score is directionless in the sense that its deviation from zero bears no relationship to the direction of underlying voxel activation in the region. That is, if all voxels in a region are more active in condition one than condition two in half the subjects, but the reverse pattern holds in the remaining subjects, all subjects will yield positive classification summary scores for that region. This score can then be submitted to a group t-test in order to test the null hypothesis that, in the population of subjects, the region's activity pattern does not differ robustly across the experiment conditions.

This difference between the directional summary scores of the GLM and the directionless summary scores of MVPA is crucial for my point, and is illustrated in Figure 4.1. The directional GLM score exposes within-subject confounds to counterbalancing in the group test, whereas the directionless MVPA score does not. To see this, consider an experiment with 10 participants and two conditions, and suppose that these conditions have no real effect on BOLD signal in any brain area whereas condition presentation order has a robust, positive effect in region X. Suppose further that presentation order is duly counterbalanced across participants, with five participants experiencing condition A first, and five participants experiencing condition B first. The order effects would drive region X, so that the first > second contrast yielded significant activation, but the conditions *per se* would have no effect, so that with a properly counterbalanced design, the A > B contrast yielded no activation. Now

[2]Here regions contain multiple voxels. Regions may be functionally or anatomically defined ROIs, or "searchlights" defined as small spheres of voxels throughout the brain (Kriegeskorte et al., 2006).

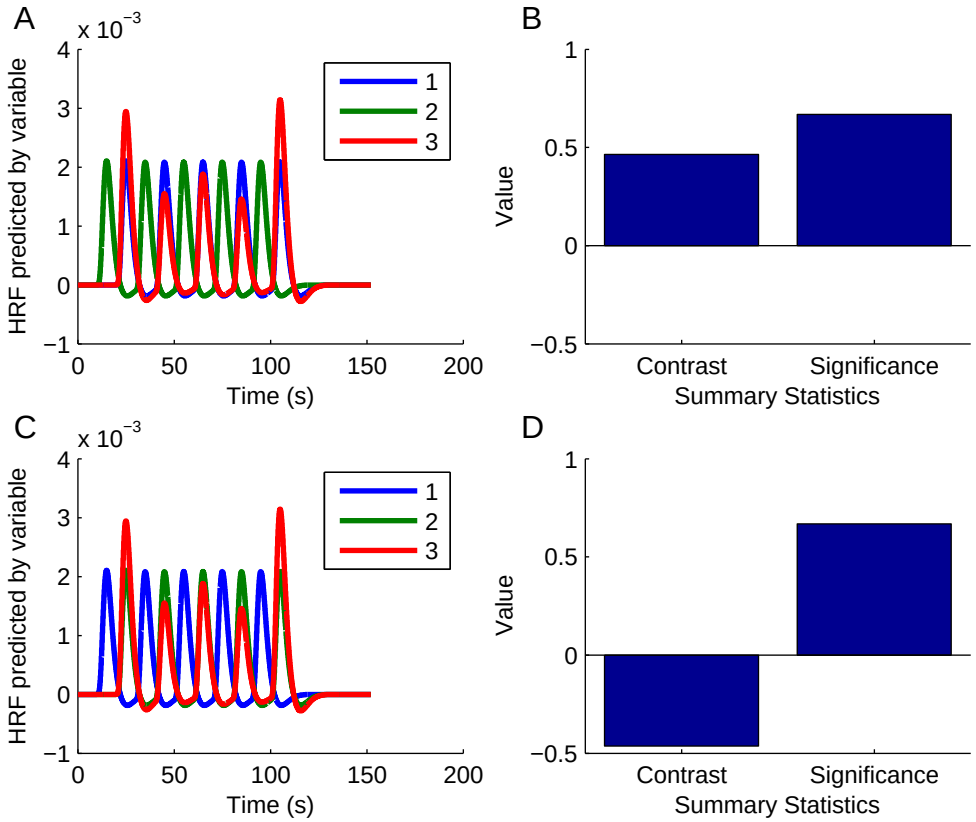



Figure 4.1: Across-subject counterbalancing is effective in the GLM, but not in MVPA. Variables 1 and 2 are experiment conditions. Variable 3 is confounded with experiment condition, and counterbalanced across subjects. (A) Predicted hemodynamic response to variables 1, 2, and 3 in first subject. (C) Predicted hemodynamic response to variables 1, 2, and 3 in second subject. (B,D) Results of single-subject GLM and MVPA analyses discriminating between variables 1 and 2, in region that is driven only by 3. (B) First subject's GLM summary score (1>2 contrast) and MVPA summary score (z-scored significance) are both positive in this region. (D) Second subject's GLM summary score (1>2 contrast) and MVPA summary score (z-scored significance) are negative and positive, respectively. In the GLM group t-test, the contrast scores will be opposite-signed across subjects and will cancel out, maintaining integrity of counterbalanced design. In contrast, in the MVPA group t-test, the z-scored single-subject significance scores will be same-signed (positive) across subjects and will not cancel out, invalidating across-subject counterbalancing design.

consider using the GLM versus MVPA to analyze these experimental data.

In the GLM, all single-subject maps would exhibit contrast scores in region X indicating effects of condition A versus B, due to the order effect confound. However, half of the subjects would exhibit positive A > B contrast scores, whereas the other half would exhibit negative A > B contrast scores. This would lead to a group mean of approximately zero for these voxels, and the group t-test would fail to reject the null hypothesis. That is, integrity of the counterbalanced design would be preserved, as expected and intended.

Similarly to the GLM, with MVPA all single-subject maps would exhibit classification scores in region X indicating effects of condition A versus B, due to the order confound. Critically, unlike with the GLM, with MVPA all subjects' scores would go in the same direction, namely positive, indicating within-subject robustness irrespective of the underlying direction of voxel activation. Since all single-subject scores were positive, the group t-test would reject the null hypothesis, reporting a significant effect of experiment condition, when in fact the order effects alone had driven the BOLD signal. Here MVPA would have undermined the integrity of the counterbalanced design.

Clearly, this would lead to the profile of MVPA success and GLM failure, and perhaps to the conclusion that region X contained a distributed representation of of the A > B difference. In fact, the result would have arisen from a simple confound. How can this situation be avoided? One solution is to specify more specific hypotheses, leading to more direction-sensitive summary statistics. RSA (Kriegeskorte et al., 2008a) offers a newer variation on MVPA mapping that, in some cases, seems to move beyond this confound by testing more directional hypotheses. In RSA, the single-subject score at each voxel reflects the degree to which the pairwise correlations between patterns across more than two experiment conditions reflect a hypothesized

relationship.[3] This hypothesized relationship can be specified by a contrast, similar to the GLM, and the resulting directionality can eliminate confounds in some but not all cases. For example, in an experiment with conditions A, B, C, D, and E, if it were hypothesized that the pairwise correlations between condition patterns should be ranked as AB>AC>AD>AE, and there was in fact no effect of condition but there was an order effect, then no spurious positive results would emerge. This follows from the fact that A would be presented close to E for some subjects, but close to B for others, so that the correlations AB and AE would be identical in the group, inconsistent with the hypothesized contrast. However, RSA would not remove other confounds. For example, consider an experiment in which conditions A, B, C, D, and E indicated compound rules composed of simple rules in such a way that the number of simple rules in common varied parametrically among condition pairs, so that A and B shared three simple rules, A and C shared two simple rules, A and D shared one simple rule, A and E shared no simple rules. For the analyst attempting to map simple rule representations with RSA, this would lead the to the same hypothesized ranking of pairwise correlations as given above. However, suppose that some of the simple rules were more difficult than others. This difficulty confound would lead to the hypothesized pattern of pairwise correlations, and to a spurious conclusion of successful rule mapping. That is, regions that responded to difficulty would be misinterpreted as responding to rule *per se*. Thus RSA MVPA mapping can eliminate confounds in some, but not cases.

Other than more directional summary scores, another possible solution, and one that has been used to augment the distributed representation argument (e.g., Cole et al., 2011), is to spatially de-mean within each anatomical region. That is, one can substract the across-voxel mean of each example used in classification. At first, this appears to be an effective control against accidentally interpreting patterns of

[3]Note that, when there are just two experiment conditions and thus two patterns, RSA reduces to basic, classification-based MVPA mapping.

broad co-activation as distributed patterns. Therefore, it appears to provide control for confounds that would lead to co-activation patterns, such as difficulty or order effects. However, it is unlikely that a co-activation effect will be perfectly homogeneous within a voxel region that was defined *a priori*, based either on voxel geometry, as in searchlight mapping analyses (Kriegeskorte et al., 2006), or on widely agreed upon anatomical subdivisions, such as V1 (Kamitani & Tong, 2005). Moreover, even in regions defined functionally by co-activation, there is likely to be inhomogeneity in the strength of activation across voxels within the region. Therefore, after de-meaning, classifier examples are still likely to be informative with regard to confounding variables, and spuriously informative with regard to variables of interest, albeit less so. I empirically demonstrate the ineffectiveness of this measure in the case study presented below.

Instead, I propose the removal of potential confounding effects from the fMRI dataset during preprocessing by simple linear regression. Specifically, I propose a two-pass process. First, covariates of no interest can be removed from the entire voxel time series (e.g., in a simultaneous regression with a basis set used for high-pass filtering). Then, after the the voxels are collected into a set of samples for the MVPA classifier, covariates of no interest can again be removed from the entire classifier sample in order to capture additional linear relationships between confounds and voxels that may have been slightly nonlinear in the original fMRI timeseries.

One issue with this proposal is that care must be taken not to saturate the regression models. That is, the number of specified covariates should be small relative to the number of TRs in the first pass, and small relative to the number of classifier examples in the second pass. This is particularly important during the second pass, as there are typically many fewer classifier examples than TRs. At issue is the fact that the degrees of freedom absorbed by the covariates are not considered in the subsequent MVPA analysis (and it is not obvious how they could be incorporated). This

could potentially lead to inflated risk of Type I error. To take an extreme example where there are 100 classifier examples and 99 covariates that are all linearly independent of each other and are orthogonal to experiment condition, the second pass would cause all voxels to be perfectly correlated with experiment condition. This would lead to spurious, perfect classification. Nonetheless, this is a problem already faced with classical GLM analyses, in which the single-subject degrees of freedom are typically discarded prior to the group t-test, and yet the removal of covariates is standard practice in that setting.

Another issue is that this method of confound control only works when there is an imperfect relationship between the confound and the variable of interest. Otherwise, removal of the confound by this method will completely remove all linear signal related to the variable of interest. That is not to say that complete removal is unwarranted, but rather that it produces the trivial effect of eliminating the variable of interest, perhaps rendering the dataset uninteresting. To illustrate, I consider a single experiment design in two different order effect confound cases. Suppose that experiment conditions A and B are ordered so that one condition tends to come before the other, and that which condition comes before the other is counterbalanced across subjects: one subject might experience more A before B (AABAABAABBABBABB), whereas another might experience more B before A (BBBABBABABAAABAA). In the first case, suppose that the order confound that drives the signal is trial number. This is imperfectly correlated with condition, and may be removed by linear regression without trivially removing all linear A and B effects. In the second case, suppose that the order confound that drives the signal is simply which condition came first. Here, the order effect is perfectly correlated with experiment condition at the single-subject level — it is a partitioning of the trials that perfectly recapitulates the A/B partitioning at the single-subject level. Therefore, in this second case, removing the relevant order effect by the proposed method would be equivalent to completely removing the

variable of interest from the dataset.

In summary, I have demonstrated that across-subject counterbalancing is ineffectual in standard, classification-based MVPA mapping. This leads to many confounds that cannot easily be removed from the experiment design. In situations where confounds must be present, linear regression is standard throughout neuroscience, psychology, and the social sciences as a method for controlling for them. I propose the use of linear regression in a two-pass procedure for the control of these confounds in MVPA. Although this procedure is not without issues, I propose it as a powerful, and perhaps necessary tool, that should be used judiciously. I turn now to a case study, demonstrating in a novel rule representation mapping dataset how confounds that are controlled in GLM analyses can contaminate MVPA results, and demonstrating the use of this proposed method of control.

## 4.3 Experimental Demonstration: Rule Representations

Flexible rules are stimulus-response mappings that are typically novel to the subject, that must be maintained in working memory, and that change frequently according to external demands. Mapping these representations in healthy human adults has been a central goal central of fMRI research on cognitive control and working memory. Although classical GLM methods have been able to map which brain areas are more active when rule representations are required, these methods have not proven sensitive enough to conclusively ascertain where the rules *per se* are represented. That is, GLM methods can typically detect areas that are differentially activated when some rule is required as opposed to baseline rest, but generally cannot detect areas that are differentially activated when one rule as opposed to another is required.

With reported MVPA mapping successes in other representational domains demon-

strating the greater sensitivity of these newer methods, there has been a recent surge of MVPA rule representation mapping studies. For example, in one early study, Haynes et al. (2007) reported that the intention to perform addition versus subtraction was represented in medial frontal cortex. Bode and Haynes (2009) later reported a striking within-trial temporal progression of rule representations from parietal to prefrontal cortex. More recently, Reverberi et al. (2011), Woolgar et al. (2011), and Cole et al. (2011) have all reported similar success. I began a similar line of MVPA-based rule mapping. However, I stopped in order to revise the approach upon recognizing the confound issue described here. I present my initial experiment results here, for the purpose of demonstrating the powerfully misleading effect of single-subject confounds within MVPA mapping.

To anticipate, the confound that I identify and control for appears to be between rule condition and individual differences in difficulty, perhaps due to individual differences in how the different rules are learned or encoded. That is, some subjects may find one rule more difficult, whereas others may find the other rule more difficult. Since the identity of the more difficult rule varies randomly across participants, this individual difference is effectively counterbalanced, and is not reflected in either the behavioral ANOVA or the group GLM results. However, it is reflected in the MVPA classification results. Due to the fact that RT is widely used as an index of difficulty, trial-to-trial RT formed a suitable covariate for use with the proposed method of confound control.

### 4.3.1 Methods

#### Participants

Forty subjects (22 male; age M=21.11 years, SD=3.82 years) were recruited from the Princeton University community for Experiment 1. Experiment sessions were approximately 1.5 hours long, and participants received $20/hour plus a performance-based

bonus of approximately $10 for participating in the experiment. The experiments were approved by the Princeton University Institutional Review Board for human subjects research. Equipment failure introduced artifacts that compromised one subject's fMRI data. This subject was discarded from all analyses.

### Task Design & Procedure

The task was a simple three-alternative forced choice design, similar to that used by Bode and Haynes (2009). Subjects used fingers of the right hand to respond using left, right, and down buttons on a response box. Trials were either cue trials or probe trials. On cue trials, participants saw a cue stimulus (one of the characters 'A' or 'B') instructing them with regard to the subsequent trials, and they pressed Down to acknowledge the cue. On probe trials, participants saw one probe stimulus (one of the characters 'X' or 'Y'), and responded according to the rule set by the most recently seen cue. The rules were: if A, then X=Right, Y=Left; if B, then X=Left, Y=right. Thus the two rules had to be maintained in working memory, and required incompatible or incongruent stimulus-response mappings, forcing participants to flexibly working memory contents whenever a cue appeared.

Each trial began with a 0.5s stimulus presentation. Participants had 3s to respond. On trials where participants failed to respond within 3s, a feedback screen appeared for 1s with a red background and black text reading "TIMEOUT. Please respond faster." On error trials, a feedback screen appeared for 0.5s with a red background and black text reading "INCORRECT RESPONSE." There was no immediate feedback on correct trials. After the response and feedback, a blank screen was presented for the remainder of the intertrial interval (ITI), which was 10+/-0.5s from stimulus onset to stimulus onset. The temporal jitter was continuous and uniform. There were 54 trials per block, and five experiment blocks.

Each trial had a 50% chance of being either a new cue (i.e., a rule switch) or a

probe stimulus, with the constraint that at least one probe always followed a cue. Each probe stimulus had a 50% chance of being either 'X' or 'Y'. Therefore, rule, probe, and response were statistically balanced within-subject.

Participants received practice with the an accelerated version of the experimental paradigm (1s ITI, no jitter) outside of the scanner. They practiced until reaching 90% accuracy on 96 consecutive trials. This typically required less than 10 minutes of practice. Participants who could not achieve 90% accuracy within 20 minutes of practice were paid for their practice time and dismissed from the remaining experiment. Participants who successfully reached criterion went on to complete the five experiment blocks inside the scanner. They were then paid for their time based on the length of the session and their performance.

### FMRI data acquisition

Subjects were scanned using a 3T Siemens Allegra head-only MRI scanner. Structural images were acquired using a T1-weighted MPRAGE (field of view, 256mm; matrix, 256x256; 176 1mm sagittal slices). Functional images were acquired using a T2*-weighted echo-planar pulse sequence (repetition time, 2s; echo time, 30ms; flip angle, 75 degrees; in-plane resolution, 3x3mm; slice thickness, 3mm; 1mm interslice gap; interleaved slices; slices aligned to anterior-posterior commissural axis). Each of the five experimental blocks corresponded to an echo-planar functional acquisition run. Due to the jittered ITI, runs were variable in length, but each included approximately 270 volumes.

### fMRI Analysis

Data were processed using SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK) and custom code written in Matlab (The MathWorks, Inc.). During preprocessing, functional (T2*) volumes were realigned to the

first volume to correct for head motion, using a rigid-body transformation which also provided head motion parameters. The mean realigned functional volume was then registered to the anatomical (T1) volume using an affine transformation. The anatomical volume was then transformed to Montreal Neurological Institute (MNI) standard space using a high-dimensional nonlinear warp. For the GLM, the warp parameters were applied to the functional volumes, which were then spatially smoothed with an 8mm FWHM Gaussian kernel. For MVPA analyses, each volume was spatially normalized (i.e., z-scored across voxels).

I were interested in rule representations, so the analyses described below were conducted on the difference between rule A and rule B. To most clearly demonstrate the confound with RT (a proxy for difficulty), I analyzed the BOLD response to probe onsets. I analyzed all trials together, as well as rule switch trials in isolation, since overall RT is longest on these trials (Monsell, 2003). To demonstrate that the effect of RT confound control was specific to the rule mapping analyses, I conducted an additional analysis in which I mapped the difference between left and right responses to probes. Error trials were excluded from all analyses.

I used the GLM as implemented in SPM8. This included a 128s high-pass filter. For rule analysis, I included in the design matrix two event-related regressors, one for each rule, time-locked to probe onset. For the response analysis, I included two event-related regressors, one for each response, time-locked to probe onset. In both analyses, I also included a single event-related regressor that was time-locked to cue onset (i.e., one regressor was shared by both 'A' and 'B' cues, in order to capture the generic cue response). Regressors were convolved with the canonical hemodynamic response function included with SPM8. I included the six head motion parameters that were estimated during preprocessing as covariates of no interest. After estimating the model, single-subject A > B contrast maps were entered into a group t-test.

I used searchlight MVPA (Kriegeskorte et al., 2006) with 4-voxel radius search-

lights. This afforded exploratory, wholebrain mapping, and is similar to analyses reported in recent rule representation articles (Bode & Haynes, 2009; Reverberi et al., 2011; Woolgar et al., 2011; Cole et al., 2011). I used the Gaussian Naïve Bayes (GNB) classifier (Pereira, Mitchell, & Botvinick, 2009). The unit of sampling for the classifier was single-trial BOLD signal. That is, the BOLD response at 6s after each probe onset was estimated from the two surrounding TRs using linear interpolation, and then the estimates were labeled by condition (i.e., A or B for the rule analysis, left or right for the response analysis) and grouped by run into training and testing sets. I generated single-subject summary scores using leave-1-run-out cross-validation over the five scanning runs (corresponding to the five blocks of the behavioral experiment). Specifically, the summary score for each subject and each searchlight was the number of correctly classified test examples, summing across all cross-validation folds, minus half the total number of test examples. Thus these scores were symmetrically distributed about zero under the null hypothesis of no representation in that searchlight. Single-subject summary score maps were then spatially transformed into MNI standard space and entered into group t-tests.

As proposed, I conducted an additional preprocessing step during MVPA analysis in which I removed covariates of no interest. I conducted two types of MVPA analyses, "typical" and "extended," that differed in which covariates were removed. I used a two-pass removal process. First, covariates were removed from the fMRI time series by linear regression, simultaneously with low frequency confounds (i.e., filtering). This was done separately for each of the five scanning runs. Second, covariates were removed from the classifier examples (training and test sets were combined for this purpose), to capture any linear relationship between covariates and classifier examples that had been slightly nonlinear in the original fMRI time series. For the first pass, and for both typical and extended analyses, the design matrix including a 20s finite impulse response model of generic cue trials as well as six head motion parameters

and a discrete cosine basis corresponding to a high-pass filter with 128s cutoff period. Thus this was similar to the covariates included in GLM analyses. In the extended analyses only, I additionally removed trial-to-trial RT in both the first and second pass. In the typical analysis, no covariates were removed from the second pass. Thus the typical and extended analyses were identical, except that RT was removed only in the extended analysis.

I also explored the effect of subtracting off the mean of each MVPA searchlight, a practice intended to control confounds and further support distributed pattern interpretations. I conducted this mean subtraction in an additional version of the typical A vs B MVPA analyses.

Group-level t-tests results were conducted and corrected for wholebrain cluster-based significance at the 0.05 level using threshold-free cluster enhancement (TFCE Smith & Nichols, 2009), a parameter-free correction algorithm.

### 4.3.2 Results

#### Behavior Results

For probe trial RT, a rule (A, B), by response (left, right), by trial type (rule switch vs rule repeat) within-subjects ANOVA showed a main effect of response, so that left was slower than right, $F(1,38)=12.87$, $p=0.0009$; and a main effect of trial type, with switch trials slower than repeat trials, $F(1,38)=12.02$, $p=0.0013$. No other effects reached significance. Critically, there was no main effect of task, $p=0.5743$, indicating that any task RT differences tended to wash out in the group.

Given the specific process by which confounds can contaminate MVPA results, it is relevant to consider the relationship between rule and RT on a single-subject basis. For this, I conducted separate two-sample, two-tailed t-tests for each of the 39 subjects. This revealed 13 subjects with significantly different RT across the two rules. Importantly, eight subjects were significantly slower in the A rule than the

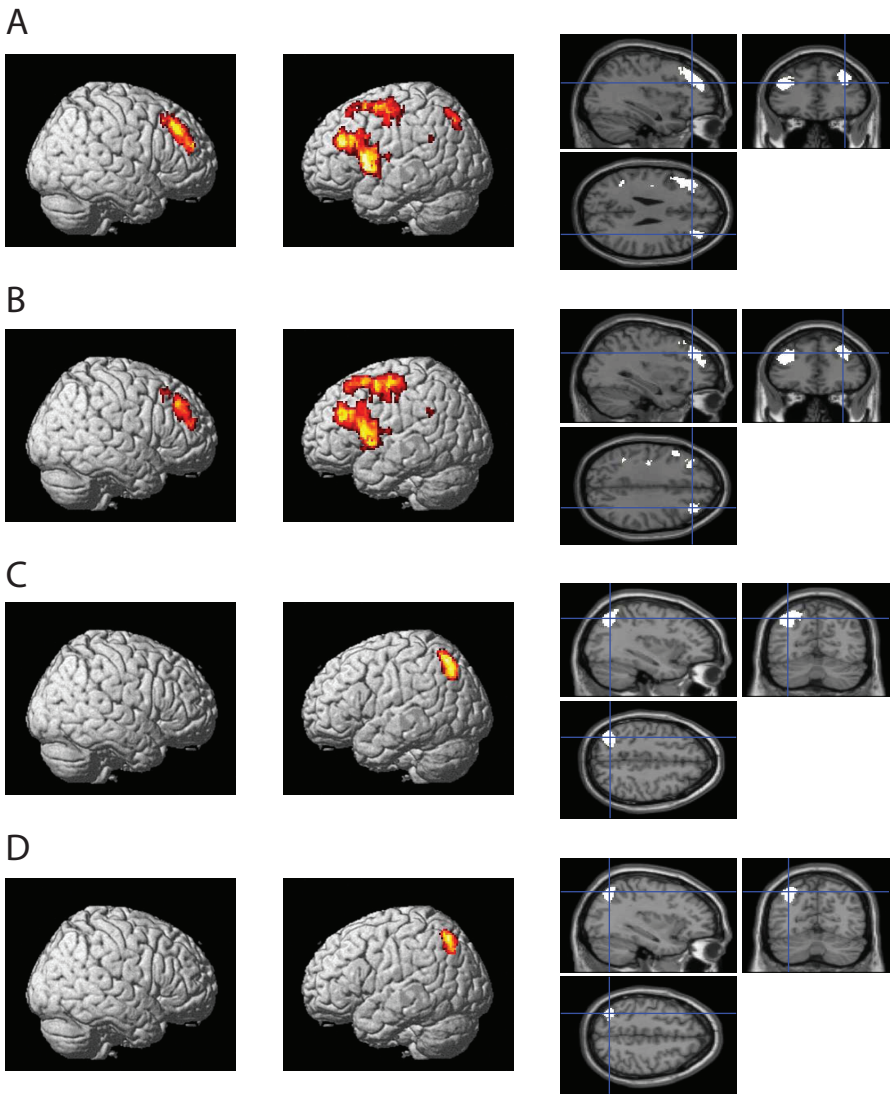



Figure 4.2: Typical MVPA analysis (i.e., not controlling for RT) appears to reveal robust rule representation maps. (A) Switch-trial-only analysis. (B) Switch-trial-only analysis, subtracting the mean across voxels from each searchlight. (C) All-trial analysis. (D) All-trial analysis, subtracting the mean across voxels from each searchlight. Maps show one minus wholebrain corrected p-values, thresholded at 1-p=0.95. Cluster-based correction was conducted using TFCE (Smith & Nichols, 2009).

B rule, whereas five subjects exhibited the opposite significant effect. I can use a binomial test to further ask whether this eight-to-five split is different than would be expected by chance if the population produced A > B subjects and A < B subjects with equal probability. This test fails to reject the null hypothesis, b(13,0.5)=8, p=0.1134, consistent with the idea that this observed pattern of single-subject, RT-rule relationships reflects a random individual difference that will tend to wash out in directional analyses such as ANOVA and the GLM, but that may confound MVPA. In other words, the pattern of behavior results suggests that the RT-rule relationship in this dataset is useful for demonstrating my thesis.

### 4.3.3 fMRI Results

As expected, the GLM revealed no significant differences in the A > B contrast. However, as shown Figure 4.2, the typical MVPA analysis appeared to reveal robust differences in several regions that are largely consistent with recent rule representation mapping reports. These included bilateral middle frontal gyrus, left pre- and post-central gyrus, and left caudate in the switch-trial analyses, and and left superior parietal lobule in both switch-trial and all-trial analyses. Moreover, as seen in the figure, there was virtually no effect of subtracting the searchlight mean from each classifier example, with the only qualitative difference being that superior parietal lobule was no longer evident in the switch-trial-only analysis. However, the extended analysis, in which RT was removed via the two-pass regression process described above, revealed no significant differences between A and B anywhere in the brain (minimum corrected p-value across all searchlights was greater than 0.2 in both the switch-trial and all-trial extended analyses).

To further establish that an RT confound drove MVPA classification results, I conducted two auxiliary GLM analyses. Results are shown in Figure 4.3. In one analysis, I included RT as a parametric regressor, and tested this regressor against

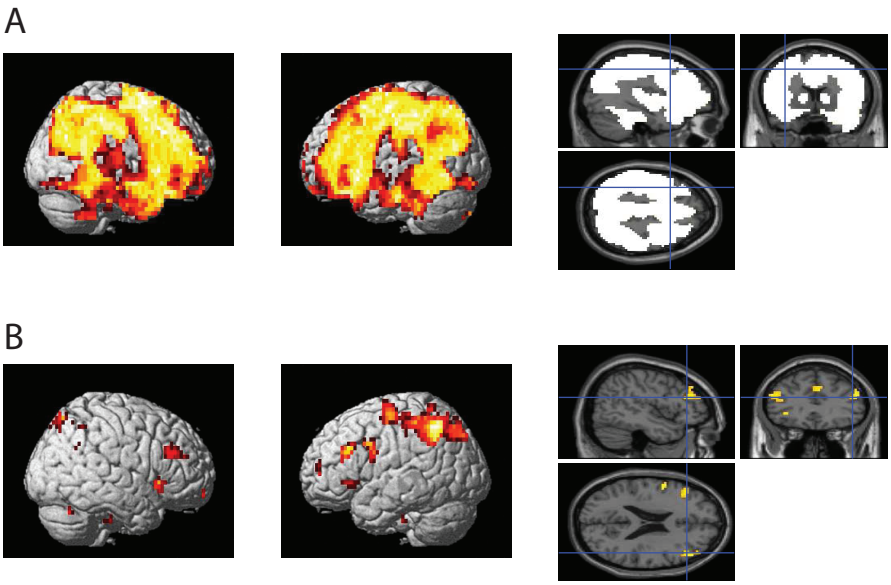



Figure 4.3: GLM analysis reveals RT effects. (A) Most cortical voxels are significantly correlated with RT. Map shows one minus wholebrain corrected p-values, thresholded at 1-p=0.95. Cluster-based correction was conducted using TFCE (Smith & Nichols, 2009). (B) Slow rule > fast rule contrast produces weak effects, reflecting the strong relationship between RT and voxels, and the weak relationship between rule and RT. Map shows group-level t-statistics, thresholded at 0.01, uncorrected, for illustration purposes.

zero in the group t-test. It is clear that RT exhibited correlations with BOLD signal throughout the brain. Indeed, over half of the voxels in the probabilistic gray-matter mask (thresholded at 0.1) included with SPM8 were found to be significantly correlated with RT after wholebrain correction. In another, I determined which rule led to faster probe RT separately for each subject, and then examined the slow rule > fast rule contrast. Given the strength and broad distribution of the RT correlation, I reasoned that even a weak relationship between rule and RT on a per-subject basis would lead to diffuse patterns of weak, positive activation in response to the slow rule > fast rule contrast. Indeed, a weak, diffuse, positive pattern is shown in 4.3B. This contrast did not produce significant activation after wholebrain correction, but it is displayed at a liberal threshold ($p<0.01$, uncorrected) for the purpose of illustrating the spatial overlap between the GLM activation and the decoding success seen in the typical MVPA analysis. It is entirely plausible that this pattern, which is weak in each voxel but consistent across many voxels, would lead to robust MVPA mapping effects, since MVPA can aggregate over weak signals in many voxels more effectively than the GLM. Indeed, a GLM ROI analysis, using the areas where MVPA rule classification was successful as a mask, confirmed that slow rule trials had significantly more activation than fast rule trials, $t(38)=1.726$, $p=0.0462$. The spatial overlap between the slow > fast GLM result and A versus B MVPA result suggests that the reason that MVPA succeeded in rule classification due to the confound with RT.

To further clarify, I note that although the GLM A > B contrast produces different results than the slow > fast contrast, the typical MVPA slow rule versus fast rule result is precisely identical to the typical MVPA A versus B result. This follows from the difference between directional GLM summary scores and directionless MVPA summary scores, and from the way slow and fast rule variables were defined. That is, the slow/fast variable was defined so that it was perfectly collinear or confounded with the A/B variable within each subject's dataset, but not at the group level. Given

the perfect single-subject confound, MVPA classification on either A versus B or fast versus slow would produce identical results — the slow versus fast information passes perfectly into the A versus B analysis, and the reverse is also true. In contrast, in the GLM, since the fast rule was A in some subjects but B in others, the slow > fast result washes out in the A > B group result. This is another way of stating my overall thesis, and of motivating these auxiliary GLM analyses in terms of determining which difference, fast versus slow or A versus B, drove the MVPA result.

I conducted an additional analysis of response direction (left or right), to show that the effect of RT confound control was specific to eliminating the rule result. Results are shown in Figure 4.4. Both the typical and extended MVPA analyses, as well as the corresponding GLM analysis, found robust differences between left and right response activity in and around left pre- and postcentral gyri, which was contralateral to the (right) hand used to make button presses. In contrast to the rule analysis, this response analysis was essentially unaffected by the removal of the difficulty confound.

### 4.3.4 Experiment Discussion

I conducted a novel rule representation fMRI study, and found results similar to those recently reported (Bode & Haynes, 2009; Reverberi et al., 2011; Woolgar et al., 2011; Cole et al., 2011). In particular, MVPA appeared to succeed in detecting regions whose activity differentiated between specific rules, whereas the GLM failed, a profile which is often interpreted as evidence for distributed representations. However, I demonstrated the presence of a single-subject confound between rule condition and difficulty, for which RT was a proxy. This confound was eliminated in the normal stages of GLM analysis but was admitted by MVPA, and it accounted for the success of the MVPA mapping analysis, undermining the conclusion that the map reflected rule representations *per se*. Although the method of removing the mean of each

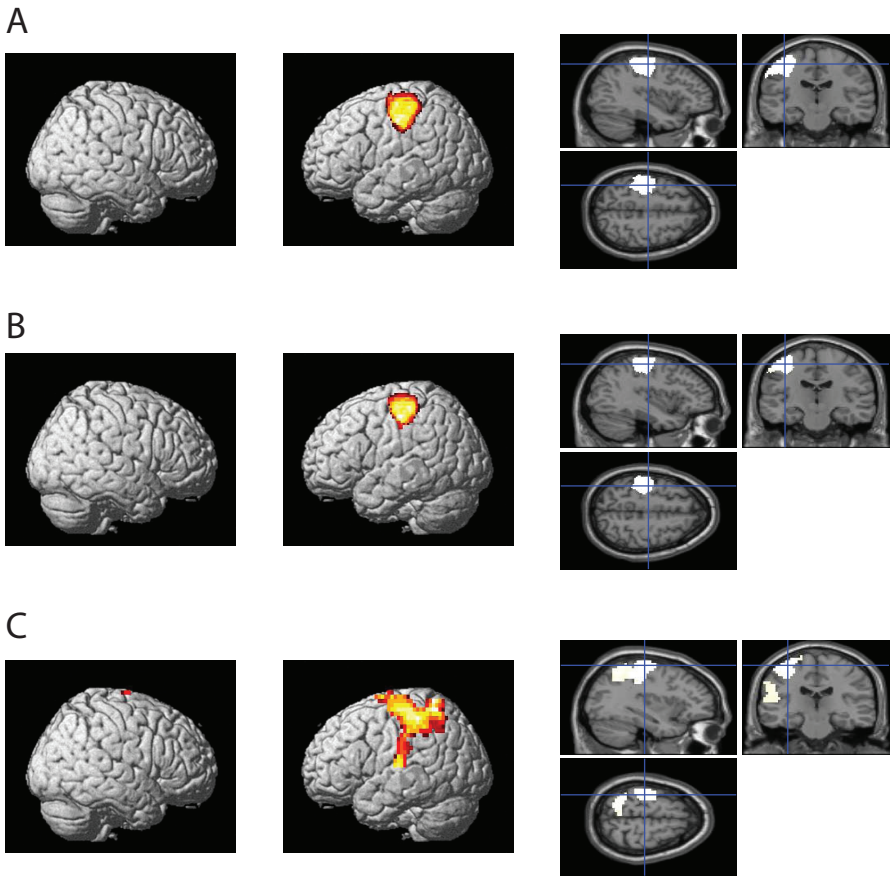



Figure 4.4: Response direction (left versus right) exhibited robust differences in both MVPA and GLM after controlling for RT. (A) Typical MVPA analysis. (B) extended MVPA analysis. (C) GLM analysis. Maps shows one minus wholebrain corrected p-values, thresholded at 1-p=0.95. Cluster-based correction was conducted using TFCE (Smith & Nichols, 2009).

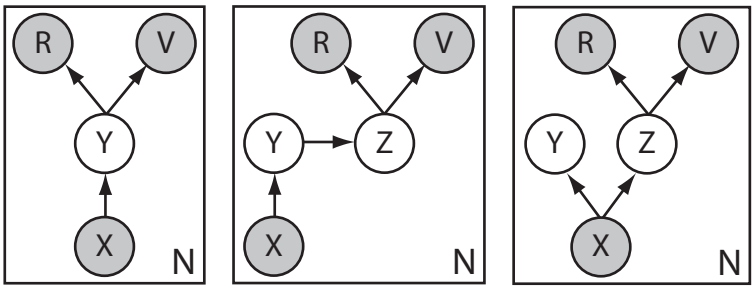



Figure 4.5: Graphs depicting different causal structures that may have given rise to the observed data. X denotes experiment condition, Y denotes rule representation, R denotes RT, V denotes voxels, and Z denotes some unspecified mental state (perhaps mental effort). N denotes number of trials. Graphing convention follows Bishop (2007). Shaded and unshaded notes denote observed and unobserved variables, respectively.

searchlight (Cole et al., 2011) left the initial results essentially unchanged, removing RT by a two-pass linear regression process reduced the rule mapping results to chance. A second analysis mapping response direction (left or right), succeeded with both MVPA and the GLM, and was unaffected by the removal of RT, indicating that the effect was specific to the rule analysis. This was a particularly stringent specificity test, since RT and response direction were in fact confounded at both the single-subject and group level.

As is always the case when confounds are present, at issue is the pattern of cause and effect. As illustrated in Figure 4.5, one possibility is that rule drove (caused) an internal rule representation, which in turn caused the observed voxel patterns and RT. If this is the case, then I would have been correct to present the results as a map of rule representations. Another possibility is that rule condition drove an internal rule representation, but also drove another mental state such as difficulty, and that this other mental state, not rule representation, caused observed voxels patterns and RT. A related possibility is that rule condition drove internal rule representation, which itself drove the other mental state, which then drove RT and voxels. If either of these latter two causal structures gave rise to the data, then I were right to be

cautious and to not present the results as a map of rule representations.

The difficulty is that the experiment design does not allow us to adjudicate among these possibilities, a problem that may apply to many classification-based MVPA studies. At first, parsimony might seem to favor interpreting the results as a map of rule representations, since this requires the analyst to invoke fewer unobservable mental state variables. However, a parsimonious specification of the research hypothesis includes only experiment condition, rule representation, and voxels, without including RT at all — this must be extended *post hoc* to accommodate the link between rule and RT. Moreover, RT is widely used as a measurable proxy or index of task difficulty, and confounds between different rule (or working memory) conditions and difficulty are well established in the GLM-based human neuroimaging literature, where additional design controls are used to address them (e.g., D'Esposito et al., 1995; Gevins, Smith, McEvoy, & Yu, 1997; Barch et al., 1997). These design elements will presumably need to be adapted and extended for use in future MVPA rule mapping studies.

Although I do not present it in detail here, I conducted a follow up experiment with 40 new subjects. The design was similar to that described above, but the response deadline was reduced from three seconds to one second. I reasoned that this would reduce the confound with difficulty by reducing RT variability, while preserving the differences between rules. MVPA on this follow up dataset did not identify rule representations anywhere in the brain, although a response direction analysis produced results that were nearly identical to those reported above. This result is consistent with the interpretation that the initial, positive MVPA results in the above study reflected the confound with difficulty (and RT) rather than rule representations.

## 4.4 General Discussion

I have articulated a process by which confounds can lead to misinterpretation of results in MVPA studies. This problem affects classification-based MVPA mapping studies in general, but effects RSA MVPA mapping studies only in some cases, and may affect other applications of MVPA very little. The confounds enter due to the inherently directionless single-subject summary statistics that classification-based MVPA mapping employs. Unlike the directional single-subject summary scores produced by the GLM, the MVPA scores shield within-subject confounds from across-subject counterbalancing that is designed to take effect during group tests. This can lead to spurious results, due to variables that are commonly counterbalanced across subjects, such as order effects, and due to individual difference variables, such as which condition is more preferred or more difficult, that are not manipulated by the experimenter but that vary randomly across subjects. These confounds provide an alternative explanation for the profile of success with MVPA and failure with the GLM that is usually interpreted as evidence for distributed representations.

I presented, as a case study, a novel rule representation fMRI experiment, similar to others recently reported in the literature (Bode & Haynes, 2009; Reverberi et al., 2011; Woolgar et al., 2011; Cole et al., 2011). I showed that a confound between RT and rule condition drove MVPA mapping success. Given the ambiguity of cause-and-effect relationships introduced by the confound, conservative interpretation of these results required that I not conclude that the results reflected rule representations rather than some other variable such as task difficulty.

Although confounds should ideally be removed during experiment design, it may prove difficult to address confounds, such as order effects, that are typically counterbalanced across subjects but that cannot be counterbalanced in this way with MVPA. I therefore proposed a method for controlling these confounds by a two-step linear regression process. Linear regression is a widely used tool for removing linear effects of

otherwise-unavoidable confounds, and I showed empirically that it was both effective and specific. In contrast, I argued and showed empirically that the method of subtracting the mean from each classifier example (e.g., Cole et al., 2011) was ineffectual for confound control.

I noted that RSA (Kriegeskorte et al., 2008a), a newer variant on standard MVPA mapping analyses, may be useful for controlling for within-subject confounds in some cases. This works because representational similarity analyses require data analysts to specify more specific, directional hypotheses. However, I noted this is not a panacea, and confounds must still be considered on a case-by-case basis.

I focused on this article specifically on MVPA mapping. Another major branch of MVPA application is so-called mind-reading analysis. In these applications, analysts extract a scalar signal from wholebrain voxel patterns, and then relate this signal to behavior in order to adjudicate among competing behavioral models. The claims made in these types of MVPA studies may often be unaffected by the confound problem I described. Specifically, whenever the claims depend only on the relationship between *some* extracted signal and behavior without depending on the identity of the signal, they are unaffected by this confound issue. For example, Haynes et al. (2007) reported that a signal can be extracted from fMRI that indicates subjects' intentions several seconds prior to the time at which subjects behaviorally report awareness of their own decision. Here, it is irrelevant which variables actually drove the extracted signal — it is sufficient that *any* brain signal predicted the choice before subjects themselves reported it. However, the line between mapping and mind-reading applications can blur. In the same paper, Haynes et al. (2007) used a searchlight analysis to locate the source of this early information. They concluded that the intention (i.e., to perform one or the other of two arithmetic operations) was represented in medial prefrontal cortex. This conclusion is problematic due to the confounds described in this paper, because there could be many variables that were confounded

with arithmetic operation. Therefore, with regard to MVPA mind-reading studies, the confound issue presented here must be considered on a case-by-case basis.

In conclusion, MVPA is a powerful new set of tools for the analysis of fMRI data. However, these tools are still evolving, and I have demonstrated one way in which they systematically admit confounds that can lead to overinterpretation of MVPA mapping results. Newer forms of MVPA mapping, such as RSA, may help to reduce this problem. For standard, classification-based MVPA mapping methods, I have proposed a solution that seems straightforward, effective, and specific, and is applicable retroactively to many previously collected datasets.

# Chapter 5

# Increasing Sensitivity and Interpretability of fMRI Mapping with Model-Based Multivoxel Pattern Analysis

## 5.1 Introduction

In the previous chapter, I argued that fMRI mapping of context representations is crucial for testing aspects of GAT and Gating models, yet success in this domain has proved elusive. The issue may be that these representations simply cannot be mapped using data from current fMRI instruments. Alternatively, it may be that context representation maps can be produced with existing datasets, but existing statistical techniques are poorly suited for this. Indeed, I noted that existing MVPA mapping methods introduce a confound that obscures interpretation of existing context representation maps, and that eliminating this confound by a fairly crude method also eliminates the maps. It may be that using less confounded or more sensitive

MVPA methods would restore these maps or produce different ones. In this chapter, I propose a new family of model-based MVPA methods that theoretically increase sensitivity and reduce confounds in mapping analyses. After fully describing the new family of methods, I apply them to the data described in the previous chapter. Unfortunately, the new methods did not change the null results reported in the previous chapter. However, theory and simulations described below suggest that the model-based MVPA approach is promising, and warrants further development.

### 5.1.1 Model-based MVPA

I refer to the family of methods developed here as model-based MVPA to denote that the main concept is to incorporate cognitive models into classification based MVPA mapping. This has two effects. First, it increases hypothesis specificity or directionality over standard classification-based MVPA mapping, thereby reducing the confounds described in the previous chapter. Second, it increases sensitivity by allowing additional data, such as trial-to-trial RT fluctuations, to inform classifier parameter estimates. This incorporation of cognitive model into MVPA is achieved by adapting a generative classifier — one that depends upon a "forward model" of the data generating process — to include elements of a particular cognitive model. This generally requires the inclusion of an unobservable or *latent* cognitive state variable, and the specification of dependencies between this and observable variables such as voxels, behavior, and experiment condition.

The theoretical benefit of the model-based MVPA approach arises from the fact that specifying the cognitive model as part of the MVPA classifier yields a more specific sensitivity profile. That is, a more specific generative model yields reduced Type II error in brain regions where the model's assumptions are met, and reduces sensitivity in situations where the model's assumptions are not met. In contrast, standard generative classifiers can be thought of as using multi-purpose generative models that

afford a lower level of sensitivity in a broad range of situations. This is another way of framing the directional/directionless distinction raised in the previous chapter. Model-based MVPA creates a sensitivity peak where neural activity relates to both experiment condition and behavior in a hypothesized way. In contrast, standard MVPA is generically sensitive when neural activity relates to experiment condition, irrespective of the relationship with behavior. These claims are made explicit in simulations below.

The model-based MVPA approach offers a general template for creating MVPA classifiers that are tailored to particular cognitive domains. Thus, when standard MVPA mapping fails (or is ambiguous due to confounds) and a cognitive model is available for the domain of interest, model-based MVPA provides an avenue for further exploration of the data. In this chapter, I give an overview of the template, and focus on the specific use of GAT for model-based MVPA mapping of context representations.

In the remainder of this introduction, I first describe the generative classifier approach as well as the particular generative model used in the standard Gaussian Naïve Bayes (GNB) generative classifier. I then discuss how GAT (Cohen et al., 1990; Miller & Cohen, 2001) may be used to adapt this standard generative model for context representation mapping.

### 5.1.2 Generative classifiers and model-based MVPA

Generative MVPA classifiers share a characteristic procedure. Given experiment condition labels for a sample of observations or examples across a set of voxels: (1) separate the sample into training and test sets; (2) fit the forward or generative model to the training set; (3) invert the fitted generative model to form a discriminative model; (4) use the discriminative model to classify each example in the test set; (5) compute inferential tests by comparing test set classification to actual labels.

This procedure differs from discriminative classifiers only in steps (2) and (3). In discriminative classifiers, these are replaced with a single step, which is to directly estimate the parameters of a discriminative model, such as the support vector machine. Generative and discriminative approaches each have strength and weaknesses. However, the generative approach most readily supports model-based MVPA, because the generative model is inherently modular, and is a natural place to introduce a domain-specific cognitive model.

The GNB is the standard generative MVPA classifier. It is defined by its generative model, which is a multivariate Gaussian with a diagonal covariance matrix for each experiment condition. Indeed, the words "Gaussian" and "Naïve" in the name refer, respectively, to the multivariate Gaussian distribution and the diagonal covariance matrix, whereas the word "Bayesian" simply means "generative."

When the GNB is used in MVPA, its generative model (Figure 5.2) embodies cognitive assumptions. Although this fact is usually overlooked, it is the central technical motivation for model-based MVPA. That is, the first step in developing model-based MVPA for a particular domain is to enumerate the implied cognitive assumptions and ask whether they are appropriate. For example, if one believes that voxels are Gaussian given cognitive state rather than Gaussian given experiment condition *per se*, then the GNB's generative model implies that experiment condition and cognitive state are equivalent. While this assumption is likely to be violated in most datasets, I argue below that it may be particularly inappropriate in the domain of context representations. Due to the modularity of the generative model, this assumption can be changed. In general then, one may identify and then change inappropriate components of the generative model, yielding a new classifier, with different properties. When incorrect assumptions are replaced with assumptions that match the true data generating process, the resulting classifier has reduced risk of Type II error. When specific assumptions are added, the resulting classifier is less

sensitive to confounds that drive neural activity but violate the assumptions. These points form the motivation of the model-based MVPA approach.

### 5.1.3 Fluctuations in Task Representations

In the domain of context representations, the GNB assumption that experiment condition is equal to cognitive state may be substantially incorrect. Instead, cognitive state — context representations — may fluctuate relative to experiment condition. The reason is that context representations are working memory representations, meaning that they require active maintenance rather than simply being driven by a persistent external stimulus. The MVPA literature includes many examples of how memory representations in general fluctuate relative to a fixed experiment condition (e.g., Polyn et al., 2005; McDuff, Frankel, & Norman, 2009; Quamme, Weiss, & Norman, 2010; Chiu, Esterman, Gmeindl, & Yantis, 2011). Such experiments usually employ a "study-test" structure, in which participants are first exposed to all experiment conditions (e.g., different item categories) in a study phase, and are then exposed to a neutral condition (e.g., a generic recall cue) in a test phase. In the study phase, each experiment condition is presented for a contiguous block of time, and is assumed to reliably drive cognitive state within that block. A classifier is thus trained on the study phase data, where equivalence between labels and cognitive states presumably holds. The trained classifier is then applied to the test phase, in which cognitive state fluctuates relative to the (fixed) experiment condition. The classifier's output is then related to fluctuations in behavior over time in the test phase. In general, classifiers can successfully decode the cognitive state during the test phase, as shown by good fits between classifier output and test phase behavioral fluctuations. Moreover, classifier output can be used to highlight the fact that behavior itself fluctuates, following underlying fluctuations in cognitive state.

Given that standard MVPA classifiers succeed in measuring other kinds of fluc-

tuating memory representations in study-test experiment designs, why should these same methods fail to measure fluctuating context representations. One possible answer is that, in context representation experiments, it is not obvious how to construct a study phase analogous to that used in other memory experiments. Instead, context representations by definition require that participants rapidly switch among tasks (i.e., experiment conditions). This precludes the contiguous block design typically used in the study phase, and perhaps leads to trial-to-trial fluctuations in the match or mismatch between cognitive state and experiment condition. For example, the task switching literature provides clear evidence for fluctuations in the degree to which context representations are actually instantiated (e.g., Monsell, 2003). In contrast, if participants were allowed to perform one task repeatedly (as in a task "study" phase), learning mechanisms might begin to make the unchanging stimulus-response mappings at least partly automatic, obviating context representations *per se* (Cohen et al., 1990; Miller & Cohen, 2001). As a result, in context representation experiments, the sample of measurements from each experiment condition must be assumed to reflect a mixture of multiple cognitive states. This implies that parameter estimates for the GNB (and also discriminative MVPA classifiers, which implicitly make similar assumptions) may be degraded, due to accidental averaging over samples from multiple context representations. Moreover, this type of degradation problem would lead specifically to inflated risk of Type II error. This problem is illustrated in Fig 5.1A-D.

### 5.1.4 GAT as a Generative Model of Context Fluctuations

Given that fluctuations may inflate the risk of Type II error when using standard MVPA classifiers to measure context representations, we can address this by incorporating the fluctuations into the generative model, forming a model-based MVPA classifier. The first step is to adapt the generative model of the GNB so that experi-

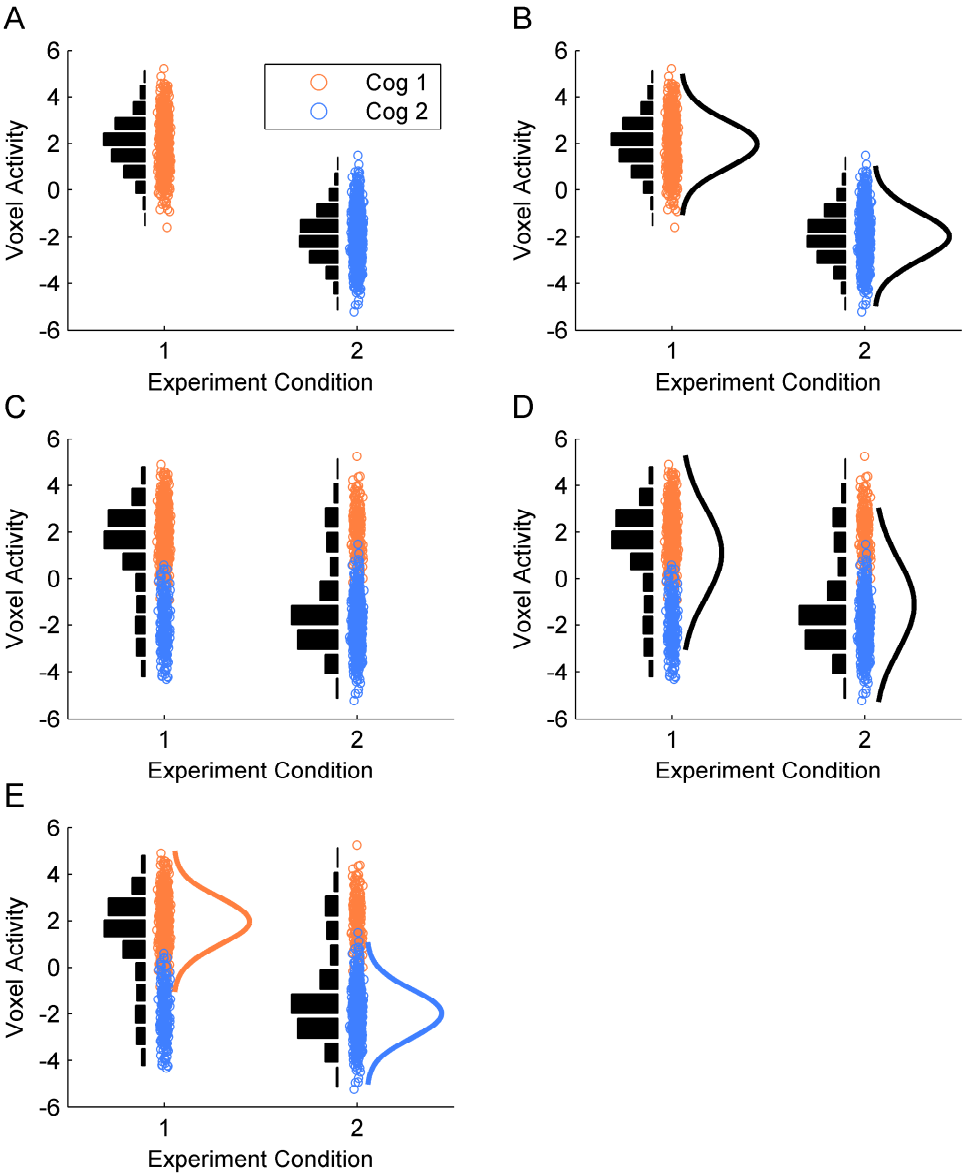



Figure 5.1: Standard MVPA classifiers assume that experiment condition (x-axis) and internal cognitive state (scatter plot colors) are perfectly correlated. (A) Assumption met: each experiment condition's sample contains measurements from only one cognitive state. (B) Standard classifiers yield accurate estimates of parameters for each cognitive state. (C) Assumption violated: each experiment condition's sample contains a mixture of measurements from multiple cognitive states. (D) Standard classifiers yield inaccurate estimates of parameters for each cognitive state; the estimated difference between means decreases while variance increases. (E) A non-standard classifier that assumes that cognitive state fluctuates. This classifier will rely less on experiment condition labels, recovering accurate parameter estimates and revealing the separability of activation across cognitive states.

ment condition and cognitive state are separate variables. Such a classifier could, in principle, recover accurate parameter estimates for each cognitive state, reducing the risk of Type II error. This is shown in Fig 5.1E. The second step is to specify generative model components that describe how context fluctuations relate to experiment condition and behavior. For this, I used aspects of GAT. GAT states that context representations enable appropriate stimulus-response mappings. This is achieved by first selectively facilitating processing of appropriate stimulus dimensions in more posterior brain regions. This selective facilitation ultimately biases motor processing in such a way as to promote context-appropriate behavioral responses. GAT is typically implemented in connectionist models (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Braver & Cohen, 1999; Brown, Reynolds, & Braver, 2007; Cohen et al., 1990; Miller & Cohen, 2001; Rougier et al., 2005), which emphasize competition among parallel representations at each processing level (e.g., among context representations, among stimulus dimension processing regions, and among motor responses), and excitatory, often bidirectional, connections that communicate state information continually among different levels. Processing noise is important as well, particularly in the motor layer, where it produces a distribution over RT and accuracy that can be compared to real data. These noisy parallel competitions translate into a model of RT and accuracy behavior. In particular, trial-to-trial variability in RT and accuracy can be explained by fluctuations in context representations. Importantly, these connectionist GAT implementations have been used almost exclusively for modeling behavioral data, and have not been used as generative models of fMRI voxel data.

Since GAT is usually implemented in connectionist models that lack fMRI voxel components, incorporating it into a generative model for use in MVPA is challenging. Difficulties arise because, as a statistical application, classification-based MVPA mapping imposes stringent requirements on the generative model. First, the model's context representation layer must be fit to actual voxel data, substantially increasing

model dimensionality and thus the computational demands of parameter estimation relative to typical connectionist applications. Second, in a generative classifier the model must be invertible, meaning that the conditional distribution over experiment conditions given voxels and behavior must be computable.[1] In contrast, connectionist GAT models typically do not have to be inverted, and inverting such a model is a non-trivial challenge. Third, mapping requires applying the model to many different brain areas in order to generate an image of those brain areas where the model fits well. In the most widely used and arguably most unbiased type of "searchlight" mapping (Kriegeskorte et al., 2006), the model is fit to the local neighborhood surrounding every voxel in the brain. This implies that the generative model must be estimated and inverted approximately 50,000 times per subject, per cross-validation fold. This means that parameter estimation and model inversion must be computationally efficient enough to be performed millions of times per analysis. Although mapping imposes strong efficiency constraints on the generative model, it is important here for two reasons. First, a central goal in measuring context representations is testing the hypothesis that they reside specifically in dorsolateral PFC, and this requires a spatially unbiased approach such as searchlight analysis. Second, the model of fluctuating context representations may plausibly be valid in local neighborhoods of voxels, but not in the whole brain. Thus the entire approach depends on being able to isolate local voxel neighborhoods in which fluctuations result from a single representation, which in turn suggests that the model must be applied separately in many different local neighborhoods, rather than once in the whole brain.

Given the demands of MVPA and the complexity of connectionist GAT models, it was practical to implement a fairly restricted approximation of GAT in this first

[1]The model must be inverted during cross-validation. This inversion step is obviated if the generative model is fit to data for similarity, rather than classification, analyses (Kriegeskorte, Mur, & Bandettini, 2008b). However, as described later in this chapter, inversions are still required as part of the Expectation-Maximization algorithm for model fitting. Therefore, it is usually important to use a generative model that can be inverted.

model-based MVPA formulation. The key factor in finding an appropriate approximation was the tradeoff between fidelity to the connectionist versions and simplicity. Fidelity makes the results more interpretable in terms of cognitive theory, whereas simplicity makes the model more computationally efficient. I present different model variants in this chapter, noting where applicable the trade-off between fidelity to the theory and practical tractability. In general, the core GAT-like element that I include is the link between context representation fluctuations and behavior, such that when cognitive state and experiment condition match, RT is fast and accuracy is high, but when cognitive state and experiment condition mismatch, RT is slow and accuracy is low. It is relatively straightforward to incorporate variants of this basic element into a generative model that is suitable for use in an MVPA classifier.

## 5.2 Modeling Introduction: The GNB Classifier

As noted above, I will present different variants of a GAT-inspired model-based MVPA. Each is a generative classifier, meaning that it is entirely described by the generative model specification, the parameter estimation procedure, and the model inversion procedure. It is instructive to begin with the details of the GNB itself, because it is the standard MVPA generative classifier, and will serve as a useful point of reference. I therefore describe the GNB in some detail in this section, before moving on to describe model-based MVPA variants in subsequent sections.

### 5.2.1 Generative Model Specification

The GNB's generative model for a dataset with $N$ measurements is depicted graphically in Figure 5.2. The likelihood, or joint distribution over observable variables

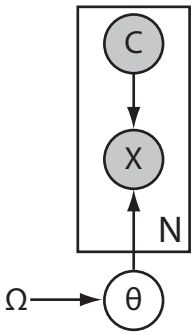



Figure 5.2: Graphical depiction of generative model from GNB, following graphing convention in Bishop (2007). Variable definitions given in main text.

(voxels and experiment condition) given parameters, is:

$$\begin{aligned} p(\mathbf{X}, \mathbf{C}|\theta) &= p(\mathbf{X}|\mathbf{C}, \theta)p(\mathbf{C}), \\ &= \prod_{n=1}^{N} p(x_n|c_n, \theta)p(c_n), \end{aligned}$$

where $x_n = (x_{n,1}, x_{n,2}, ..., x_{n,D})^T$ is a $D$-vector of voxel activity at time $n$, $c_n = (c_{n,1}, c_{n,2}, ..., c_{n,K})^T$ is a multinomial (indicator) $K$-vector giving the experiment condition at time $n$, and $\theta = (\mu, \sigma^2)$ contains the mean and variance parameters for Gaussians for each experiment condition. For notational compactness, $\mathbf{X}$ and $\mathbf{C}$ denote $(x_1, x_2, ..., x_N)$ and $(c_1, c_2, ..., c_N)$, respectively. For a single time point (omitting the $n$ subscript for simplicity), the probability distribution over voxels given experiment condition and parameters is a D-dimensional Gaussian with diagonal covariance matrix:

$$\begin{aligned} p(x|c, \theta) &= \mathcal{N}_D\left(x|\mu_c, \mathrm{diag}(\sigma_c^2)\right) \\ &= \prod_{d=1}^{D} \mathcal{N}\left(x|\mu_{cd}, \sigma_{cd}^2\right), \\ p(c) &= 1/K. \end{aligned}$$

where diag($a$) is a square matrix with the elements of vector $a$ along its main diagonal and zeros elsewhere, and $K$ is the number of experiment conditions. Note that, for this classifier application, the probability distribution over experiment conditions is discrete uniform. That is, this distribution is determined by the number of experiment conditions in the application, and does not include free parameters. This is true for all models in this chapter.

The conjugate prior distribution over parameters is the Normal-Gamma distribution (e.g., Bishop, 2007) with hyperparameters $\Omega = (\mu_0, \kappa_0, \alpha_0, \beta_0)$. That is,

$$\begin{aligned} p(\theta|\Omega) &= \prod_{k=1}^{K} p(\mu_k, \sigma_k^2|\Omega) \\ &= \prod_{k=1}^{K}\prod_{d=1}^{D} p(\mu_{kd}, \sigma_{kd}^2|\Omega) \\ &= \prod_{k=1}^{K}\prod_{d=1}^{D} \mathrm{NG}(\mu_{kd}, \sigma_{kd}^2|\mu_0, \kappa_0, \alpha_0, \beta_0) \end{aligned}$$

### 5.2.2 Parameter Estimation

Throughout this chapter, I use maximum likelihood estimation (MLE) or, more generally, maximum a posteriori (MAP) to estimate model parameters. Although the reader is likely to be familiar with these concepts as well as the concept of using logarithms to simplify maximization, I introduce them somewhat thoroughly here because parameter estimation plays an important part in the ultimate form of model-based MVPA methods. I introduce MLE and MAP here in the context of estimating GNB parameters.

MLE and MAP estimate statistical parameters by solving optimization problems. MLE accepts a model and dataset and then finds model parameters that maximize

the dataset's likelihood. For the GNB, we have:

$$\theta_{\text{MLE}} = \arg\max_{\theta}\{p(\mathbf{X}, \mathbf{C}|\theta)\}.$$

Similarly, MAP accepts a model, dataset, and prior distribution, and finds model parameters with maximum posterior probability given the dataset and priors. MAP for the GNB is:

$$\theta_{\text{MAP}} = \arg\max_{\theta}\{p(\theta|\mathbf{X}, \mathbf{C}, \Omega)\} = \arg\max_{\theta}\{p(\mathbf{X}, \mathbf{C}|\theta)p(\theta|\Omega)\}.$$

MLE and MAP have very different theoretical properties, but MLE estimates can usually be obtained from MAP estimates by a certain setting of hyperparameters, $\Omega$. Therefore, the remainder of the chapter will emphasize MAP results, from which MLE results follow as a special case.

The maximization operation suggests that we must find the roots of the gradient of the posterior with respect to the parameters:[2]

$$\nabla_\theta \left(p(\mathbf{X}|\mathbf{C}, \theta)p(\theta|\Omega)\right) = \nabla_\theta \left(\prod_{n=1}^{N} p(x_n|c_n, \theta)p(\theta|\Omega)\right).$$

Due to the product over time points, this appears to be a difficult computation. However, for any positive function $f(x)$, including the posterior distribution, we have $\arg\max_x\{f(x)\} = \arg\max_x\{\log f(x)\}$. Note also that $\log\prod_i x_i = \sum_i \log x_i$. Thus it suffices to maximize the log posterior, meaning that we need only compute the

[2]The fact that there are conjugate priors for the GNB implies that the MAP estimate solutions are readily available in reference texts. However, it is instructive to approach this simple problem ingenuously, as the solution strategy will prove useful for more difficult problems encountered below.

gradient of the log posterior, which has sums rather than products:

$$\begin{aligned}\nabla_\theta \log p(\mathbf{X}|\mathbf{C},\theta) + \nabla_\theta \log p(\theta|\Omega) &= \nabla_\theta \left( \sum_{n=1}^{N} \log p(x_n|c_n,\theta) + \log p(\theta|\Omega) \right), \\ &= \sum_{n=1}^{N} \nabla_\theta \log p(x_n|c_n,\theta) + \nabla_\theta \log p(\theta|\Omega).\end{aligned}$$

Using this log trick on the simple generative model of the GNB leads to the following well known MAP estimates for the mean and variance of the Gaussian associated with the $k$-th experiment condition and $d$-th voxel (e.g., Gauvain & Lee, 1994):

$$\begin{aligned}\left[\mu_{\text{MAP}}\right]_k &= \mu_{N_k} \\ \left[\sigma^2_{\text{MAP}}\right]_k &= \frac{\beta_{N_k}}{\alpha_{N_k} - 1} \\ \mu_{N_k} &= \frac{\kappa_0 \mu_0 + N_k \bar{x}_k}{\kappa_0 + N_k} \\ \kappa_{N_k} &= \kappa_0 + N_k \\ \alpha_{N_k} &= \alpha_0 + \frac{N_k}{2} \\ \beta_{N_k} &= \beta_0 + \frac{1}{2} S_k + \frac{N_k \kappa_0 \left(\bar{x}_k - \mu_0\right)^2}{2\left(\kappa_0 + N_k\right)} \\ \bar{x}_k &= \frac{\sum_{n=1}^{N} c_{nk} x_n}{N_k} \\ S_k &= \sum_{n=1}^{N} c_{nk} \left(x_n - \bar{x}_k\right)^2 \\ N_k &= \sum_{n=1}^{N} c_{nk},\end{aligned}$$

where $\mu_0, \kappa_0, \alpha_0$, and $\beta_0$ are the four hyperparameters of the Normal-Gamma distribution. Note that, by setting $\kappa_0 = 0$, $\alpha_0 = 1$, and $\beta_0 = 0$, we recover the standard MLE results: $[\mu_{\text{MLE}}]_k = \bar{x}_k, [\sigma^2_{\text{MLE}}]_k = S_k / N_k$.

In summary, with the GNB's simple generative model, the log trick yields well

known, closed-form MAP estimates. As the generative models considered below become more complex, the log trick fails, making parameter estimation a potential source of difficulty and a constraint on model structure. In fact, recovering the usefulness of the log trick in more complex settings forms a theme of the parameter estimation strategies discussed below.

### 5.2.3 Model Inversion and Cross-Validation

For classification-based MVPA, once a generative model has been fit to training data, it must be inverted in order to apply it to test data. This is critical for using prediction-based cross-validation for inferential tests. For the GNB, once we have estimated the likelihood, $p(x|c,\theta)p(c)$, we must compute the conditional distribution over experiment conditions given voxels, $p(c|x,\theta)$. This inversion operation is a straightforward application of probability theory. With the GNB, it is also computationally simple. Recalling that experiment condition is a $K$-dimensional multinomial vector with discrete uniform prior, we have:

$$\begin{aligned} p(c_n|x_n,\theta) &= Q^{-1}p(x_n,c_n|\theta) \\ &= Q^{-1}p(x_n|c_n,\theta)p(c_n) \\ Q &= \sum_{c_n} p(x_n|c_n,\theta)p(c_n). \end{aligned}$$

Note that the normalizing constant, $Q$, is a sum over the $K$ levels of the experiment condition variable. This inversion is computationally simple because experiment condition is discrete. If experiment condition were continuous instead, the sum in the denominator would become an integral. This integral would only be easy to compute for certain models (e.g., it would be simple for Gaussian experiment conditions). Thus, inversion difficulty is another source of constraints on model structure whenever cross-validation is required.

Once this inversion is performed, it is straightforward to test the null hypothesis that the classifier and data are unrelated. One useful null hypothesis is: given the voxel data for a test example, the model is as likely to assign greatest probability to the correct experiment condition as it is to assign greatest probability to the incorrect experiment condition. That is, given that the true experiment condition is $j$, the null hypothesis says that the random variable $p(c_{nj}|x_n, \theta) - p(c_{nk}|x_n, \theta)$ is symmetrically distributed about zero. To test this, it suffices to count the number of test examples for which the classifier assigns the greatest probability to the correct experiment condition. Under this null hypothesis, this count statistic will be binomially distributed with parameters $M$ and $1/K$, where $M$ is the total number of test examples and $K$ is the number of experiment conditions. If the test statistic is large enough that the tail distribution function falls below a critical probability, say 0.05, then the null hypothesis is rejected. Note that there is no obvious reason to think count statistic would be significantly *smaller* than if it were distributed according to the binomial null distribution, so one-tailed tests are well justified in this application, as with an $F$-test.

As a final note on model inversion and testing, there are several variants of the null hypothesis used above, some of which would yield more sensitive tests. In formulating this null hypothesis (and others below), I sought test statistics whose sampling distributions were symmetrical about a well-defined value and as close to Gaussian as possible, as these properties facilitate parametric and non-parametric $t$-tests during group analysis, such as the threshold free cluster enhancement test (Smith & Nichols, 2009) used in the previous and next chapters. Toward this end, I sacrifice some sensitivity in choosing test statistics. For example, in the special case that there are two experiment conditions, the count test statistic described above will be symmetrically distributed about $M/2$, making it suitable for use in group level tests. In contrast, cross-validated log likelihood would be a continuous and thus more sensitive test

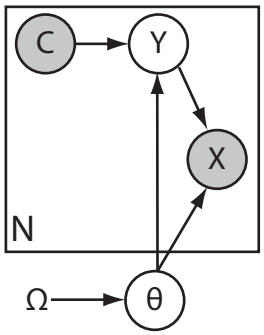



Figure 5.3: Graphical depiction of generative model variant 0, the Basic Mixture, following graphing convention in Bishop (2007). Variable definitions given in main text.

statistic, but its sampling distribution is less obviously characterized, making it less practical for use in standard group tests.

## 5.3 Model 0: Basic Mixture

Before describing the model variants including GAT-like links between RT, accuracy, and context representations, it is useful to consider first a basic model that includes fluctuating cognitive states but does not include behavior. This model is also useful empirically, as it is identical to that used in the GNB except that experiment condition and cognitive state are separate. This affords a controlled examination of the impact on classification performance of this separation, which is useful for assessing the overall feasibility of the model-based MVPA approach, as shall be shown below. Note that this newly distinct cognitive state cannot be directly measured — it is an unobservable or latent variable. This section serves as an introduction to using the Expectation-Maximization (EM: Dempster, Laird, & Rubin, 1977) algorithm, an essential tool for estimating the parameters of models that include latent variables.

### 5.3.1 Generative Model Specification

Model 0 is shown in Figure 5.3. The critical difference between this and the GNB generative model is the inclusion of a cognitive state variable, $Y$, that is distinct from experiment condition. In the figure, cognitive state is unshaded, unlike experiment condition and voxel activity, indicating that it is a latent or unobservable variable. The inclusion of this latent variable is an essential feature of all model-based MVPA variants that I develop in this chapter, and has implications for parameter estimation as well as model interpretation.

The likelihood function, or joint distribution over observable variables (voxels and experiment condition), is given by:

$$\begin{aligned} p(\mathbf{X}, \mathbf{C}|\theta) &= p(\mathbf{X}|\mathbf{C}, \theta)p(\mathbf{C}), \\ &= \prod_{n=1}^{N} p(x_n|c_n, \theta)p(c_n), \\ &= \prod_{n=1}^{N} \sum_{y} p(x_n|y, \theta)p(y|c_n, \theta)p(c_n), \end{aligned}$$

where $y$ is a $J$-dimensional indicator vector ($J \geq K$) for the latent cognitive state, and $\theta = (\mu, \sigma^2, \eta)$ contains the model parameters. As with the GNB, the distribution over voxels given cognitive state and parameters is a D-dimensional Gaussian with diagonal covariance matrix:

$$\begin{aligned} p(x|y, \theta) &= \mathcal{N}_D\left(x|\mu_y, \mathrm{diag}(\sigma_y^2)\right), \\ p(y|c, \theta) &= \mathrm{Multinomial}\left(y|\pi(c)\right), \quad \pi_i(c) = \begin{cases} \eta, & c_i = 1 \\ (1-\eta)\frac{1}{J-1}, & c_i = 0 \end{cases}, \\ p(c) &= 1/K, \end{aligned}$$

where the $J$-dimensional vector $\pi(c)$ is known as the mixture proportions, and $\eta$ is

the probability that cognitive state will match experiment condition on a given trial. That is, the cognitive state either matches the experiment condition with probability $\eta$, or is drawn from a discrete uniform distribution over the remaining cognitive states with probability $1 - \eta$. In the special case that there are only two cognitive states, this distribution becomes, $p(y|c, \theta) = \eta \mathbb{I}\{y = c\} + (1 - \eta)\mathbb{I}\{y \neq c\}$. Note that the voxel distribution is now parameterized by cognitive state rather than experiment condition *per se*, an essential difference between the GNB and model-based MVPA.

The conjugate prior for the model 0 is now the product of Normal-Gamma and Beta distributions, with hyperparameters $\Omega = (\mu_0, \kappa_0, \alpha_0, \beta_0, \eta_0)$. That is:

$$\begin{aligned} p(\theta|\Omega) &= p(\eta|\Omega) \prod_{j=1}^{J} p(\mu_j, \sigma_j^2|\Omega) \\ &= p(\eta|\Omega) \prod_{j=1}^{J} \prod_{d=1}^{D} p(\mu_{jd}, \sigma_{jd}^2|\Omega) \\ &= \text{Beta}(\eta|\eta_0) \prod_{j=1}^{J} \prod_{d=1}^{D} \text{NG}(\mu_{jd}, \sigma_{jd}^2|\mu_0, \kappa_0, \alpha_0, \beta_0) \end{aligned}$$

### 5.3.2 Parameter Estimation

In this section I describe the EM algorithm, an essential tool for estimating parameters of mixture models, including model-based MVPA. EM efficiently produces approximate MAP/MLE solutions, and can be implemented for many models with only slight modifications of standard MAP/MLE results. In particular, the EM algorithm for model 0 follows from slight modifications to the basic GNB MAP results given above. I present the EM algorithm for model 0 in this section.

#### The Expectation-Maximization Algorithm

The log trick for MAP estimation fails when working with mixture models, including all model-based MVPA variants in this chapter. The EM algorithm (Dempster et al.,

1977) is an iterative method that partially recovers the usefulness of the log trick in the mixture setting, yielding efficient, approximate MAP estimates. I introduce the EM algorithm here, and proceed to the specific formulae for model 0 in the next section.

Consider the problem of MAP estimation for model 0, which augments the GNB with a latent cognitive state variable. We can attempt to use the log trick in order to compute $\theta_{\text{MAP}}$:

$$\theta_{\text{MAP}} = \arg\max_{\theta} \{\log p(\mathbf{X}, \mathbf{C}|\theta) p(\theta|\Omega)\}$$
$$\log p(\mathbf{X}, \mathbf{C}|\theta) = \sum_{n=1}^{N} \log \left[ \sum_{y} p(x_n|y, \theta) p(y|c_n, \theta) p(c_n) \right].$$

However, the likelihood function now includes a sum over the latent variable, $y$. This is a characteristic of the likelihood functions for all mixture models, and for this particular mixture model, the logarithm cannot pass through the sum. This spoils the log trick, so that there is no longer an obvious MAP solution for model 0, even though its components remain simple Gaussians. Another way of seeing the difficulty is to consider the latent cognitive state variable as providing the labels that are used to estimate the Gaussian parameters, and then to recognize that the labels themselves must be estimated as well, because they are not observed. It is intuitive then that it is difficult to simultaneously estimate the labels and the parameters of the voxels that are Gaussian *given* the labels.

Intuitively, one might approach this chicken and egg problem by simply taking a best guess of the labels, estimating the parameters given the guess, then forming new guesses based on the parameters, and so on until convergence. This is the concept of the EM algorithm. Mathematically, this recovers some of the usefulness of the log trick, with the result that we can often apply a simple generalization to standard, closed form MAP solutions in order to achieve closed-form solutions for

each iteration of the EM algorithm. This often increases efficiency relative to other iterative optimization methods. The EM algorithm employs an auxiliary function, $F(q, \theta)$:

$$
\begin{aligned}
F(q,\theta) &= \log p(\theta|\mathbf{X},\mathbf{C},\Omega) - \sum_{n=1}^{N}\sum_{y_n} q(y_n)\log\frac{q(y_n)}{p(y_n|x_n,c_n,\theta)} \\
&= \log p(\theta|\mathbf{X},\mathbf{C},\Omega) - \sum_{\mathbf{Y}} q(\mathbf{Y})\log\frac{q(\mathbf{Y})}{p(\mathbf{Y}|\mathbf{X},\mathbf{C},\theta)} \qquad (5.1) \\
&= \sum_{\mathbf{Y}} q(\mathbf{Y})\log p(\mathbf{X},\mathbf{C},\mathbf{Y}|\theta) + \log p(\theta|\Omega) \\
&\quad - \sum_{\mathbf{Y}} q(\mathbf{Y})\log q(\mathbf{Y}) - \log p(\mathbf{X},\mathbf{C}|\Omega) \qquad (5.2) \\
&= \sum_{n=1}^{N}\sum_{y_n} q(y_n)\log p(x_n,c_n,y_n|\theta) + \log p(\theta|\Omega) \\
&\quad - \sum_{\mathbf{Y}} q(\mathbf{Y})\log q(\mathbf{Y}) - \log p(\mathbf{X},\mathbf{C}|\Omega)
\end{aligned}
$$

The EM algorithm iterates between maximizing (5.1) with respect to $q$ while holding $\theta$ constant in an "E-step," and then maximizing (5.2) with respect to $\theta$ while holding $q$ constant in an "M-step." For the E-step, note that the rightmost term in (5.1) is the negative Kullback-Leibler divergence between distributions $q(\mathbf{Y})$ and $p(\mathbf{Y}|\mathbf{X},\mathbf{C},\theta)$. This quantity is either negative or is exactly zero when the two distributions are equivalent. This implies that we can maximize $F(q,\theta)$ with respect to $q$ simply by setting $q(\mathbf{Y}) = p(\mathbf{Y}|\mathbf{X},\mathbf{C},\theta)$, or equivalently by setting $q(y_n) = p(y_n|x_n,c_n,\theta)$ for all $n$. For the M-step, note that the two leftmost terms in (5.2) are the "expected complete-data log posterior" which, taking advantaged of the current label guesses, can be maximized with respect to $\theta$ (holding $q$ constant) using computations nearly identical to MAP estimates for a non-mixture model. Thus iterating between E- and M-steps produces a sequence of parameter estimates over which the log posterior increases until reaching a local maximum. Conceptually, this amounts to iterating

between estimating the labels (E-step) and estimating the parameters given fixed labels (M-step). The benefit of this approach is that both the E- and M-steps require only well known, closed form computations. Note that this is an approximate MAP solution, and the resulting parameters may not be the global maximum.

The EM algorithm is introduced above for a mixture model with a discrete latent variable. It works equally well for continuous models by replacing sums over the latent variable with integrals. This introduces additional computational constraints on model structure. Specifically, $p(y_n|x_n, c_n, \theta)$ includes an integral that must be computed in every iteration of the E-step, and the expected complete-data log posterior becomes an integral that must be maximized in each M-step. These integrals may be difficult to work compute for models that fall outside certain structural classes.

EM often finds MAP and MLE solutions more efficiently than general purpose optimization algorithms, such as gradient ascent methods. This is understandable, given that EM was specially designed for this special class of optimization problems. General purpose methods often require first and second derivatives of the cost function (the negative posterior in this case) with respect to model parameters in order to maximize performance. These quantities can be difficult to compute for some models, whereas the EM algorithm requires only slight modification of standard MLE and MAP results. Moreover, general purpose optimization algorithms often scale poorly with the number of parameters, whereas EM scales well due to the availability of analytical maximization solutions in each M step. This scalability is important for MVPA, as MVPA models typically contain hundreds or thousands of parameters (e.g., separate mean and variance parameters for every voxel in a region, or even in the whole brain). Figure 5.4 clearly demonstrates the superiority of the EM algorithm for model 0 (derived below) over a general purpose optimization algorithm applied to the same model. The general purpose algorithm was supplied with the first but not the second derivatives with respect to model parameters, which is typical in practice

given the difficulty of computing the latter. It is important to note that the general purpose algorithm used in the simulation — the interior-point algorithm as implemented in Matlab — represents state-of-the-art nonlinear optimization. It performs poorly relative to EM only because EM was designed to leverage special structure in MAP estimation. This simulation should serve as an index of the difficulty posed by parameter estimation in model-based MVPA, which is generally a high-dimensional, nonlinear optimization problem. Due to its efficiency and scalability, the EM algorithm is an important tool for model-based MVPA parameter estimation. However, as described above, the algorithm places constraints on the structure of the models that one uses. Specifically, EM requires that $p(y_n|x_n, c_n, \theta)$ be readily computable, and that the expected complete-data log posterior be readily maximizable. If a model fails to meet this criterion, then other parameter estimation procedures that are often much less computationally efficient must be employed.

**E-step**

During the E-step for model 0, we set $q(\mathbf{Y}) = p(\mathbf{Y}|\mathbf{X}, \mathbf{C}, \theta)$, as follows:

$$\begin{aligned} q(y_n) &= p(y_n|x_n, c_n, \theta) \\ &= Q^{-1} p(x_n, y_n|c_n, \theta) \\ &= Q^{-1} p(x_n|y_n, \theta) p(y_n|c_n, \theta) \\ Q &= \sum_{y_n} p(x_n|y_n, c_n, \theta) p(y_n|c_n, \theta) \end{aligned}$$

The E-step is algebraically simple, with the small computational burden of summing over $J$ levels of the latent variable $Y$.

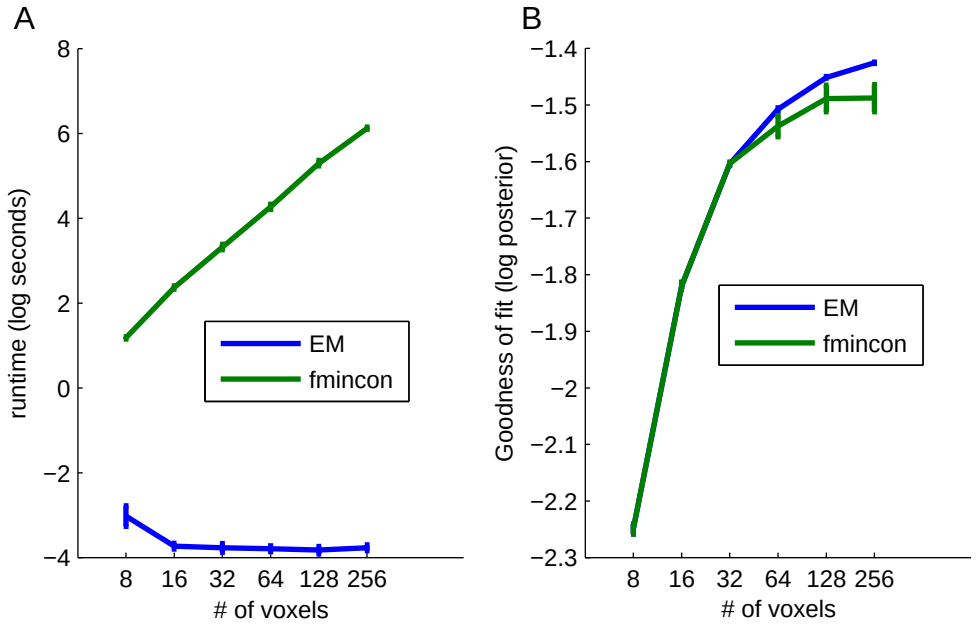



Figure 5.4: The EM algorithm outperforms a standard, general purpose optimization algorithm. (A) As the number of voxels increases to from 8 to 256, the EM algorithm consistently converges in less than 0.05s. In contrast, the convergence time of the general-purpose algorithm begins at approximately 3s for 8 voxels, and increases to 456s for 256 voxels. (B) Both algorithms initially find better parameter estimates as the number of voxels increases. However, as the number of voxels increases, the general purpose algorithm begins to find poor solutions relative to EM.

**M-step**

During the M-step for model 0, we maximize the expected complete-data log posterior, holding $q(\cdot)$ constant. That is, we must compute:

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} \left\{ \sum_{\mathbf{Y}} q(\mathbf{Y}) \log p(\theta|\mathbf{X}, \mathbf{C}, \mathbf{Y}, \Omega) \right\} \\
&= \arg\max_{\theta} \left\{ \sum_{\mathbf{Y}} q(\mathbf{Y}) \log p(\mathbf{X}, \mathbf{C}, \mathbf{Y}|\theta) + \log p(\theta|\Omega) \right\} \\
&= \arg\max_{\theta} \left\{ \sum_{n=1}^{N} \sum_{y_n} q(y_n) \log p(x_n, c_n, y_n|\theta) + \log p(\theta|\Omega) \right\} \\
&= \arg\max_{\theta} \left\{ \sum_{n=1}^{N} \sum_{y_n} q(y_n) \left[ \log p(x_n|y_n, \theta) + \log p(y_n|c_n, \theta) \right] + \log p(\theta|\Omega) \right\},
\end{aligned}$$

where I have changed to "hat" notation rather than MAP subscript to denote the fact that the M-step results do not represent final MAP estimates. By design, the M-step decouples neatly into simple maximization problems, each a straightforward extension of a well-known MAP result. For example, for the $d$-th voxel (omitting the $d$ subscript for simplicity) and the $j$-th cognitive state, we have (Gauvain & Lee,

1994):

$$
\begin{aligned}
\hat{\mu}_j &= \mu_{N_j} \\
\hat{\sigma}_j^2 &= \frac{\beta_{N_j}}{\alpha_{N_j} - 1} \\
\mu_{N_j} &= \frac{\kappa_0 \mu_0 + N_j \bar{x}_k}{\kappa_0 + N_j} \\
\kappa_{N_j} &= \kappa_0 + N_j \\
\alpha_{N_j} &= \alpha_0 + \frac{N_j}{2} \\
\beta_{N_j} &= \beta_0 + \frac{1}{2} S_j + \frac{N_j \kappa_0}{N_j + \kappa_0} \frac{(\bar{x}_j - \mu_0)^2}{2(\kappa_0 + N_j)} \\
\bar{x}_j &= \frac{\sum_{n=1}^{N} q(y_{nj}) x_n}{N_j} \\
S_j &= \sum_{n=1}^{N} q(y_{nj}) (x_n - \bar{x}_j)^2 \\
N_j &= \sum_{n=1}^{N} q(y_{nj}).
\end{aligned}
$$

This result is analogous to the GNB result given above, with the deterministic experiment condition labels replaced by the probabilistic cognitive state labels. For the latent cognitive state variable itself, we must maximize the expected complete-data log posterior with respect to $\eta$, the probability that the cognitive state and experiment condition match:

$$
\begin{aligned}
\hat{\eta} &= \frac{\eta_{N,1} - 1}{\eta_{N,1} + \eta_{N,2} - 2} \\
\eta_N &= \eta_0 + \left( \sum_{n=1}^{N} \sum_{y_n} q(y_n) \mathbb{I}\{y_n = c_n\}, \sum_{n=1}^{N} \sum_{y_n} q(y_n) \mathbb{I}\{y_n \neq c_n\} \right)^T
\end{aligned}
$$

The hyperparameter $\eta_0 = (1, 1)^T$ yields a prior for which the MAP and MLE estimates are equivalent.

Thus the M-step for model 0 is a simple generalization of the single parameter

estimation step for the GNB. The major changes are that the deterministic labels are replaced by probabilistic ones, and that we now must also estimate the parameter governing the latent cognitive state variable.

#### Log Posterior and Convergence

The EM algorithm is guaranteed to increase the log posterior until reaching a local maximum. In practice, I evaluate the log posterior after each M-step, and define convergence as occurring when the relative difference between successive log posteriors is less than 1e-6. For model 0, the log posterior is given as:

$$\begin{aligned}\log p(\theta|\mathbf{X},\mathbf{C},\Omega) &= \log p(\mathbf{X}|\mathbf{C},\theta) + \log p(\theta|\Omega) + \text{const} \\ &= \sum_{n=1}^{N} \log p(x_n|c_n,\theta) + \log p(\theta|\Omega) + \text{const} \\ &= \sum_{n=1}^{N} \log \sum_{y_n} p(x_n|y_n,\theta)p(y_n|c_n,\theta) + \log p(\theta|\Omega) + \text{const}\end{aligned}$$

Note that, in the context of the convergence check, any terms that are independent of $\theta$ can be ignored (i.e., there is no need to compute them).

### 5.3.3 Model Inversion and Cross-validation

Cross-validation with the mixture models proceeds similarly to cross-validation with the GNB. The key difference is that, as the models become more complex, it becomes important to test them more thoroughly, as shall be shown in simulations below. This can be accomplished by forming cross-validation test statistics for each of the observable variables, and then forming conjunction maps of places where all relevant test statistics reject their respective null hypotheses.

To test the null hypothesis that the model cannot predict experiment condition, one withholds the experiment condition variable for test set items, requiring the

trained model to predict it:

$$\begin{aligned} p(c_n|x_n,\theta) &= Q^{-1}p(x_n|c_n,\theta)p(c_n) \\ &= Q^{-1}\sum_{y_n} p(x_n|y_n,\theta)p(y_n|c_n,\theta)p(c_n) \\ Q &= \sum_{c_n} p(x_n|c_n,\theta)p(c_n). \end{aligned}$$

Recall that, in the context of classification, we set $p(c) = 1/K$ for all $c$, so that $p(c)$ terms cancel from the numerator and denominator. As with the GNB, to form the test statistic, I count the number of times that the model assigns greatest probability to the correct experiment condition. Under the null hypothesis, this count is binomially distributed with parameters $N$ and $1/K$, where $N$ is the number of test items and $K$ is the number of experiment conditions.

To test the null hypothesis that the model cannot predict voxels, one can withold the voxel data from each test set item, requiring the trained model to predict the mean for each voxel, as follows:

$$\begin{aligned} \mathbb{E}[x_n|c_n,\theta] &= \int_{-\infty}^{\infty} x_n p(x_n|c_n,\theta)dx_n \\ &= \sum_{y_n}\int_{-\infty}^{\infty} x_n p(x_n|y_n,\theta)dx_n p(y_n|c_n,\theta) \\ &= \sum_{y_n} \mu_{y_n} p(y_n|c_n,\theta) \end{aligned}$$

This predicted mean value can then be used as a covariate for each voxel. That is, each voxel's test set examples can be regressed onto this predicted value. The mean of the regression coefficient across voxels is symmetrically distributed around zero under the null hypothesis that the model predicts *none* of the voxels. This is the test statistic (and null hypothesis) that I use.

I have now described cross-validation test statistics for both of the model's ob-

servable variables: voxels and experiment condition. As mentioned above, more complicated models require more more thorough testing in order to interpretation results. Therefore, as shown in simulations below, a conservative test of the models presented here (other than the GNB) requires that both of these null hypotheses be rejected.

### 5.3.4 Simulations: Model 0 vs GNB

Model 0 lacks the additional assumptions about trial-to-trial behavior of the model-based MVPA classifiers described below. Nonetheless, its separation of cognitive state and experiment condition variables yields different properties from those of the GNB. In particular, Figure 5.1 suggested that a separate cognitive state variable alone should lead to detection of separable representations that would be missed by the GNB. Figure 5.5 shows simulations comparing model 0 and GNB performance when voxels fluctuate relative to or are yoked (i.e., perfectly correlated) to experiment condition. Figure 5.5A shows that the two classifiers perform identically on cross-validation tests, both when voxels fluctuate and when voxels are yoked to experiment condition. This appears to suggest that model 0 is in fact unable to compensate for fluctuations, contrary to expectations. However, the cross-validation test belies the increased sensitivity of model 0. Figure 5.5B shows that a pattern separability metric reveals that model 0 is indeed more sensitive than the GNB in the case of fluctuations, recovering essentially the same degree of separability as in the yoked case. This implies that model 0 may be a fruitful starting point for applications using MVPA generalization tests that do not rely on cross-validation, such as RSA (Kriegeskorte et al., 2008a). That is, when patterns are estimated in both train and test sets, and the generalization metric is separability of test-estimated patterns conditional on train-estimated patterns, then Figure 5.5B suggests that model 0 has a distinct advantage over the GNB. Moreover, this advantage is "free" in the sense that it does not require specification of additional assumptions about behavior or other variables. However,

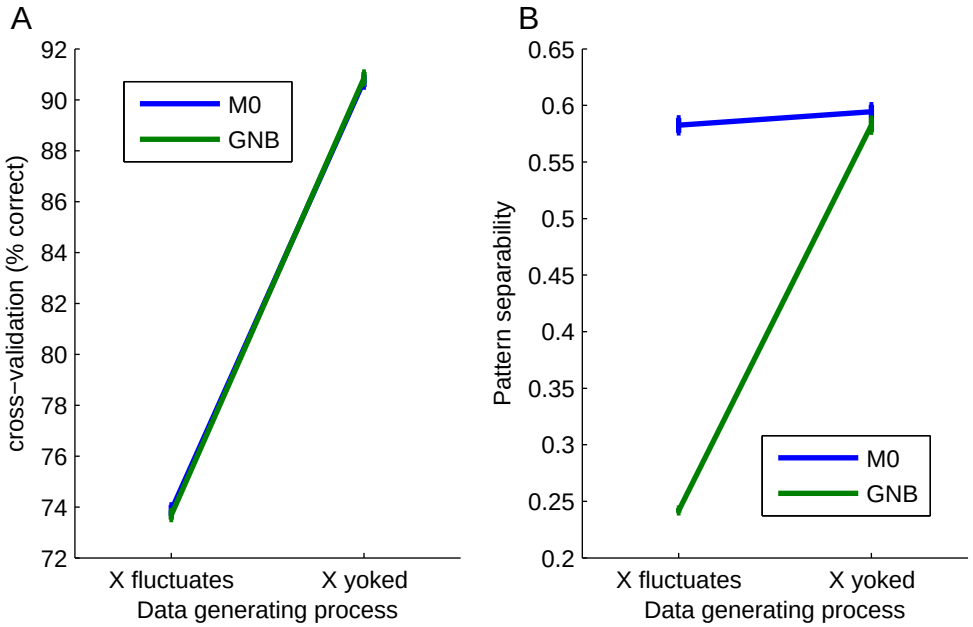



Figure 5.5: Simulation of model 0 and GNB performance across datasets where voxels (X) and experiment condition are related in different ways. (A) Cross-validation performance (50% is chance) reveals no difference between the models. (B) Pattern separability shows that model 0 is able to recover the difference between means in the where voxels fluctuate relative to experiment condition. The pattern separability metric is the mean across voxels of absolute difference between estimated means, divided by estimated variance.

this chapter focuses on the development of more specific and cross-validation based model-based MVPA methods, so potential applications to RSA are beyond the present scope.

### 5.3.5 Model 0 Discussion

Model 0 includes the latent cognitive state variable but not behavior, making it an important tool for isolating the impact of latent variables on MVPA performance. In this section I have presented the structure of model 0 together with the formulae and algorithms (including a somewhat general treatment of the EM algorithm) necessary for estimating the model's parameters from data and testing goodness-of-fit via cross-validation. Simulations suggest that model 0 yields increased sensitivity relative to

the GNB when pattern separability rather than cross-validation is used as a test metric. This has potential applications that may be developed in future work. Having introduced the basic tools for working with latent variable models, the following sections describe fully model-based MVPA models, including links between behavior and latent cognitive state inspired by GAT.

## 5.4 Model 1: Basic GAT

Model 1 is the basic version of GAT that I implemented. As with Model 0, cognitive state is a discrete latent variable that drives voxels directly. However, the model now includes behavior, which is linked in a particular way to the cognitive state (context) variable. Specifically, the match or mismatch between cognitive state and experiment condition now drives RT and accuracy. This particular type of dependency leads to a particular sensitivity profile for this model. For example, in datasets where behavior does indeed fluctuate with the match or mismatch between cognitive state and experiment condition, then in brain regions where there is a robust neural difference across experiment conditions but there is no fluctuation that matches the behavior, the model will exhibit decreased sensitivity relative to the GNB. In contrast, in brain regions where there is a weak neural signal that does fluctuate according to behavior, the behavioral measurements will facilitate detection of the representation and the model will be more sensitive than the GNB. Thus maps produced by this model-based classifier are more specific, and thus more interpretable, than maps produced by the GNB. This is demonstrated via simulations below. Here, I define the model and derive EM algorithm, inversion, and cross-validation procedures.

### 5.4.1 Generative Model Specification

Model 1 is depicted graphically in Figure 5.6.

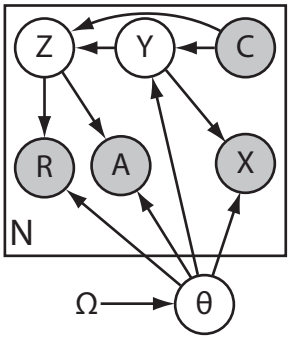



Figure 5.6: Graphical depiction of generative model 1, following graphing convention in Bishop (2007). Variable definitions given in main text.

The likelihood function is given by:

$$
\begin{aligned}
p(\mathbf{X}, \mathbf{R}, \mathbf{A}, \mathbf{C}|\theta) &= p(\mathbf{U}, \mathbf{C}|\theta) \\
&= \prod_{n=1}^{N} p(u_n, c_n|\theta) \\
&= \prod_{n=1}^{N} \sum_{y_n} p(x_n|y_n, \theta) \\
&\quad \cdot \sum_{z_n} p(r_n|z_n, \theta) p(a_n|z_n, \theta) p(z_n|y_n, c_n) p(y_n|c_n, \theta) p(c_n),
\end{aligned}
$$

where $r_n$ and $a_n$ are the RT and accuracy behavior at time $n$, respectively, and the variable $U = (X, R, A)$ is introduced to simplify notation. The variable $z_n$ denotes the match or mismatch between cognitive state and experiment condition at time $n$. All other variables are as in model 0. The parameters, $\theta = (\mu, \sigma^2, \nu, \lambda, \xi, \eta)$ enter into

the model as follows:

$$
\begin{aligned}
p(x|y,\theta) &= \mathcal{N}_D\left(x|\mu_y, \mathrm{diag}(\sigma_y^2)\right) \\
p(r|z,\theta) &= \mathrm{Gamma}(r|\nu,\lambda_z) \\
p(a|z,\theta) &= \mathrm{Bernoulli}(a|\xi_z) \\
p(y|c,\theta) &= \mathrm{Multinomial}(y|\pi(c)), \quad \pi_i(c) = \begin{cases} \eta, & c_i = 1 \\ (1-\eta)\frac{1}{J-1}, & c_i = 0 \end{cases} \\
z|y,c &= \mathbb{I}\{y=c\} \\
p(c) &= 1/K
\end{aligned}
$$

Note that the RT and accuracy distributions depend on the match or mismatch between experiment condition and cognitive state. Note also that there is a single shape parameter for the RT distribution that does not depend on the match or mismatch between cognitive state and experiment condition, whereas the rate parameter (inverse scale) does.

The prior distribution, with hyperparameters $\Omega = (\mu_0, \kappa_0, \alpha_0, \beta_0, \nu_0, \tau_0, \lambda_0, \xi_0, \eta_0)$, has the form:

$$
\begin{aligned}
p(\theta|\Omega) &= p(\eta|\Omega)\prod_{j=1}^{J} p(\mu_j, \sigma_j^2|\Omega)p(\nu|\Omega)\prod_{i=1}^{2} p(\lambda_i|\Omega)p(\xi_i|\Omega) \\
&= p(\eta|\eta_0)\prod_{j=1}^{J}\prod_{d=1}^{D} p(\mu_{jd}, \sigma_{jd}^2|\mu_0,\kappa_0,\alpha_0,\beta_0)p(\nu|\nu_0)\prod_{i=1}^{2} p(\lambda_i|\tau_0,\lambda_0)p(\xi_i|\xi_0) \\
&= \mathrm{Beta}(\eta|\eta_0)\prod_{j=1}^{J}\prod_{d=1}^{D}\mathrm{NG}(\mu_{jd},\sigma_{jd}^2|\mu_0,\kappa_0,\alpha_0,\beta_0)\mathrm{U}(\nu|0,\nu_0) \\
&\quad \cdot \prod_{i=1}^{2}\mathrm{Gamma}(\lambda_i|\tau_0,\lambda_0)\mathrm{Beta}(\xi_i|\xi_0)
\end{aligned}
$$

### 5.4.2 Parameter Estimation

In this section I describe the necessary EM algorithm for model 1. Due to the fact that voxels, RT, and accuracy are conditionally independent given the latent cognitive state, this is simply a combination of standard "mixture of Gaussian," "mixture of Gamma," and "mixture of Bernoulli" results.

#### E-Step

Recall that during the E-step, we update the probability distribution over cognitive states, $q$, holding parameters constant. For the basic GAT model, this proceeds as follows:

$$\begin{aligned} q(y_n) &= p(y_n|u_n, c_n, \theta) \\ &= Q^{-1} p(u_n|y_n, c_n, \theta) p(y_n|c_n, \theta) \\ &= Q^{-1} p(x_n|y_n, \theta) p(r_n|y_n, c_n, \theta) p(a_n|y_n, c_n, \theta) p(y_n|c_n, \theta) \\ &= Q^{-1} p(x_n|y_n, \theta) p(y_n|c_n, \theta) \sum_{z_n} p(z_n|y_n, c_n) p(r_n|z_n, \theta) p(a_n|z_n, \theta) \\ Q &= p(u_n|c_n, \theta) \end{aligned}$$

So the E-step is algebraically simple, with the computational burden of summing over $J$ discrete levels of the latent cognitive state variable.

#### M-Step

During the M-step, we maximize the expected complete-data log posterior with respect to the parameters, holding $q$ constant. The expected complete-data log posterior

is given as,

$$\begin{aligned}
\sum_{\mathbf{Y}} q(\mathbf{Y}) \log p(\theta|\mathbf{U}, \mathbf{Y}, \mathbf{C}, \Omega) &= \sum_{\mathbf{Y}} q(\mathbf{Y}) \log p(\mathbf{U}, \mathbf{Y}|\mathbf{C}, \theta) + \log p(\theta|\Omega) + \text{const} \\
&= \sum_{n=1}^{N} \sum_{y_n} q(y_n) \log p(u_n, y_n|c_n, \theta) + \log p(\theta|\Omega) + \text{const} \\
&= \sum_{n=1}^{N} \sum_{y_n} q(y_n) \log \left[ p(u_n|y_n, c_n, \theta) p(y_n|c_n, \theta) \right] \\
&\quad + \log p(\theta|\Omega) + \text{const}, \\
\log p(u_n|y_n, c_n, \theta) &= \log p(x_n|y_n, \theta) + \log p(r_n|y_n, c_n, \theta) + \log p(a_n|y_n, c_n, \theta) \\
&= \log p(x_n|y_n, \theta) + \log \left[ \sum_{z_n} p(z_n|c_n, y_n) p(r_n|z_n, \theta) \right] \\
&\quad + \log \left[ \sum_{z_n} p(z_n|c_n, y_n) p(a_n|z_n, \theta) \right]
\end{aligned}$$

Note that the sums over $z_n$ in the last step each reduce to a single term, because $Z$ is deterministic given $Y$ and $C$. Therefore, the log passes through these sums without issue.

As with model 0, the M-step for model 1 decouples neatly into optimization problems with standard closed form solutions. The mixture proportion and voxel parameter results are unchanged from those for model 0. However, we must derive the results for the RT and accuracy components. For accuracy, we have a Bernoulli variable and a Beta prior, yielding the following M-step result for the $i$-th value of $z$:

$$\begin{aligned}
\hat{\xi}_i &= \frac{f_1 - 1}{f_1 + f_2 - 2} \\
f &= \xi_{0i} + \left( \sum_{n=1}^{N} q(z_n = i) a_n, \sum_{n=1}^{N} q(z_n = i)(1 - a_n) \right)^T \\
q(z_n = 1) = \sum_{y_n} q(y_n) \mathbb{I}\left\{ y_n = c_n \right\}, &\quad q(z_n = 0) = \sum_{y_n} q(y_n) \mathbb{I}\left\{ y_n \neq c_n \right\}
\end{aligned}$$

where $\xi_{0i}$ is the 2-dimensional hyperparameter vector for the Beta prior on the $i$-th value of $z$, and $\xi_{0i} = (1,1)^T$ yields the MLE. Here $q(z_n = 1)$ is the probability under distribution $q$ that the latent cognitive state matches the experiment condition.

The M-step update for RT is slightly more complex, because there is in general no closed form MAP estimator for the shape parameter of the Gamma distribution. Given the shape parameter however, there is a conjugate prior (also a Gamma distribution) for the rate (inverse scale) parameter. I therefore assume a uniform prior on the shape and resort to 1-dimensional numerical optimization, employing the closed form solution for the rate. This yields:

$$
\begin{aligned}
\hat{\lambda}_\nu &= \left( \frac{f_1 - 1}{g_1}, \frac{f_2 - 1}{g_2} \right)^T \\
f &= \tau_0 + \left( \sum_{n=1}^{N} q(z_n = 1)\nu, \sum_{n=1}^{N} q(z_n = 0)\nu \right)^T \\
g &= \lambda_0 + \left( \sum_{n=1}^{N} q(z_n = 1) r_n, \sum_{n=1}^{N} q(z_n = 0) r_n \right)^T \\
\hat{\nu} &= \arg\max_\nu \left\{ \sum_{\mathbf{Y}} q(\mathbf{Y}) \log p(\hat{\lambda}_\nu, \nu | \mathbf{R}, \mathbf{Y}, \mathbf{C}, \Omega) \right\} \\
&= \arg\max_\nu \left\{ \sum_{\mathbf{Y}} q(\mathbf{Y}) \log p(\mathbf{R} | \mathbf{Y}, \mathbf{C}, \hat{\lambda}_\nu, \nu) + \log p(\hat{\lambda}_\nu, \nu | \tau_0, \lambda_0, \nu_0) \right\} \\
&= \arg\max_\nu \left\{ \sum_{\mathbf{Y}} q(\mathbf{Y}) \log \sum_{\mathbf{Z}} p(\mathbf{Z} | \mathbf{Y}, \mathbf{C}) p(\mathbf{R} | \mathbf{Z}, \hat{\lambda}_\nu, \nu) + \log p(\hat{\lambda}_\nu, \nu | \tau_0, \lambda_0, \nu_0) \right\}
\end{aligned}
$$

The final line specifies a 1-dimensional nonlinear optimization problem that must be solved numerically. This can be solved efficiently by widely available algorithms.

**Log Posterior and Convergence**

The log posterior is given by:

$$
\begin{aligned}
\log p(\theta|\mathbf{U},\mathbf{C},\Omega) &= \log p(\mathbf{U}|\mathbf{C},\theta) + \log p(\theta|\Omega) + Q \\
&= \sum_{n=1}^{N} \log p(u_n|c_n,\theta) + \log p(\theta|\Omega) + Q \\
&= \sum_{n=1}^{N} \log \sum_{y_n} p(u_n|y_n,c_n,\theta)p(y_n|c_n,\theta) + \log p(\theta|\Omega) + Q \\
&= \sum_{n=1}^{N} \log \sum_{y_n} p(x_n|y_n,\theta)p(r_n|y_n,c_n,\theta)p(a_n|y_n,c_n,\theta)p(y_n|c_n,\theta) \\
&\quad + \log p(\theta|\Omega) + Q \\
&= \sum_{n=1}^{N} \log \sum_{y_n} p(x_n|y_n,\theta) \\
&\quad \cdot \sum_{z_n} p(z_n|y_n,c_n)p(r_n|z_n,\theta)p(a_n|z_n,\theta)p(y_n|c_n,\theta) + \log p(\theta|\Omega) + Q,
\end{aligned}
$$

where $Q$ includes all terms that are constant with respect to $\theta$, and can be ignored.

### 5.4.3 Model Inversion and Cross-validation

Model inversion here is only slightly more complex than with model 0. Cross-validation tests are the same, except that it is necessary to additionally test RT predictions. I will not describe a cross-validation test for accuracy, as it is difficult to construct a test statistic for that variable, and RT cross-validation without accuracy cross-validation should provide sufficiently stringent null hypothesis testing.[3]

To test experiment condition prediction, we have (omitting dependence on esti-

[3]There are highly skewed numbers of accurate vs inaccurate behavior examples in most context representation experiments, and so the obvious choice of a binomial null distribution becomes problematic. That is, cross-validation statistics would be binomial with parameters $M$ (number of test examples) and $p$ (observed accuracy), and when $M$ and $p$ are, say, 100 is 0.95, respectively, the null distribution is very negatively skewed. This makes it difficult to use in group tests, which often require symmetric null distributions.

mated $\theta$ for simplicity):

$$\begin{aligned} p(c_n|u_n) &= Q^{-1} p(u_n|c_n)p(c_n) \\ &= Q^{-1} \sum_{y_n} p(u_n|y_n, c_n)p(y_n|c_n)p(c_n) \\ &= Q^{-1} \sum_{y_n} p(x_n|y_n) \sum_{z_n} p(z_n|y_n, c_n)p(r_n|z_n)p(a_n|z_n)p(y_n|c_n)p(c_n) \\ Q &= p(u_n) \end{aligned}$$

This is algebraically quite similar to, and only slightly more computationally expensive than, the experiment condition prediction for model 0. Once one has predicted experiment condition, the construction of the test statistic is identical to the same process for model 0.

To predict mean voxel activity, we can modify the result from model 0 to include behavior:

$$\begin{aligned} \mathbb{E}[x_n|r_n, a_n, c_n] &= \int_{-\infty}^{\infty} x_n p(x_n|r_n, a_n, c_n)dx_n \\ &= \sum_{y_n} \int_{-\infty}^{\infty} x_n p(x_n|y_n)dx_n p(y_n|r_n, a_n, c_n) \\ &= \sum_{y_n} \mu_{y_n} p(y_n|r_n, a_n, c_n) \\ p(y_n|r_n, a_n, c_n) &= Q^{-1} \sum_{z_n} p(r_n|z_n)p(a_n|z_n)p(z_n|y_n, c_n)p(y_n|c_n) \\ Q &= p(r_n, a_n|c_n) \end{aligned}$$

Once the voxel activity has been predicted, construction of the cross-validation test statistic proceeds as with model 0.

To test the null hypothesis that the model cannot predict RT, one withholds all behavior variables for test set items and then requires the model to predict mean RT,

as follows:

$$\begin{aligned}
\mathbb{E}[r_n|x_n, c_n] &= \int_0^\infty r_n p(r_n|x_n, c_n) dr_n \\
&= \sum_{z_n} \int_0^\infty r_n p(r_n|z_n) dr_n p(z_n|x_n, c_n) \\
&= \sum_{z_n} \frac{\nu}{\lambda_{z_n}} p(z_n|x_n, c_n), \\
p(z_n|x_n, c_n) &= \sum_{y_n} p(z_n|y_n, c_n) p(y_n|x_n, c_n) \\
&= \sum_{y_n} p(z_n|y_n, c_n) Q^{-1} p(x_n|y_n) p(y_n|c_n) \\
Q &= p(x_n|c_n)
\end{aligned}$$

As with the voxels, one can construct a test statistic by regressing test set RT onto the model's predicted mean RT. Under the null hypothesis, the regression coefficient is symmetrically distributed about zero.

### 5.4.4 Model 1 Simulations

As stated above, the purpose of model-based MVPA is to generate classifiers with specialized sensitivity profiles. Changes in sensitivity arise from reducing model misspecification, which increases sensitivity; and from adding assumptions, which reduces sensitivity in general while increasing sensitivity in hypothesized cases. Simulations in Figure 5.7 demonstrate this specialization for model 1. Classification performance was simulated when the data were generated with different dependencies between experiment condition, voxels, and RT. In particular, some datasets meet model 1 assumptions, some meet GNB assumptions, and some violate the assumptions of both models. First, the simulations verify that Model 1 does outperform the GNB when the data were generated according to its assumptions, namely that voxels and RT both fluctuate relative to experiment condition, but both according to a common

latent variable (i.e., context). Second, it is clear that both classifiers performs best when their respective assumptions are met, and that performance falls off when assumptions are violated. However, model 1 performance is more sensitive than GNB performance to assumption violations. This reflects the specificity that arises from the additional assumptions made in model 1: sensitivity is blunted when the cognitive model is incorrect. Thus model 1 results are easier to interpret than those of the GNB, as they reflect more specific hypotheses. Note that this increased specificity (or decreased robustness) is not a general property of models with latent variables, as model 0 demonstrated equivalent or superior robustness relative to the GNB. Rather, increased specificity is a function of the additional assumptions that are made.

### 5.4.5 Model 1 Discussion

Model 1 adds trial-to-trial behavior to model 0, with particular dependencies among behavior, voxels, and experiment condition that were inspired by GAT. Specifically, voxels depend on cognitive state directly, whereas behavior depends on the match or mismatch between cognitive state and experiment condition. This yields a classifier with a novel sensitivity profile that reflects aspects of GAT, demonstrating the core principle of model-based MVPA. This model is only slightly more computationally complex than model 0, although it must be tested more thoroughly in order to interpret results, as shall be shown in the Additional Simulation section below. The model is very simple relative to connectionist implementations of GAT, as it must be in order to take advantage of the efficient EM algorithm. In the next section I present a variant that includes differences between behavior across tasks as well as differences across match or mismatch between cognitive state and experiment condition.

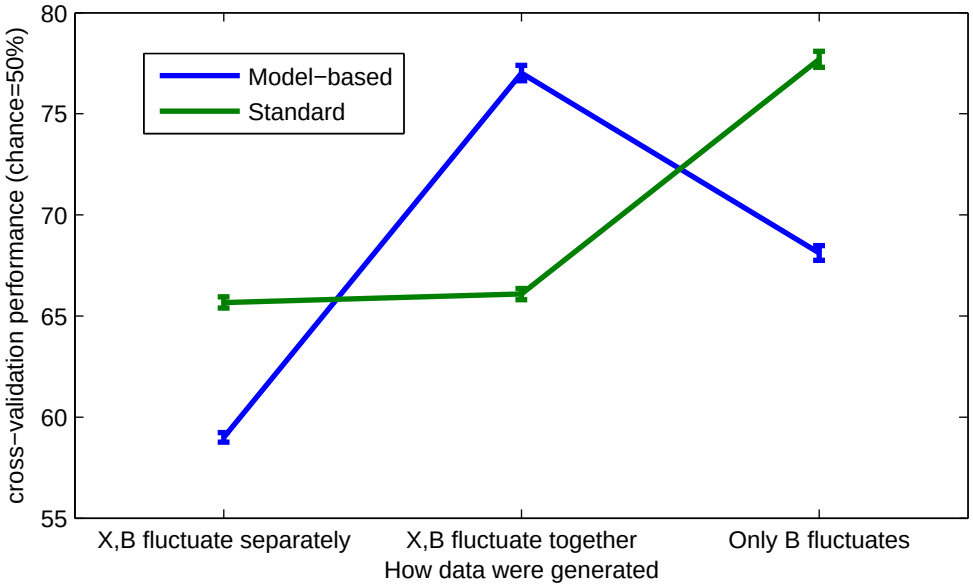



Figure 5.7: Simulation of model 1 and GNB cross-validation performance (50% is chance), across datasets where voxels (X) and RT (R) are generated in different ways. Model 1 is generally less sensitive than the GNB, except in the special case where its assumptions are met. (Left) X and R both fluctuate relative to experiment condition, but each fluctuates according to a separate latent variable. Both models' assumptions are violated. (Middle) X and R both fluctuate, and fluctuations are driven by a common latent variable (i.e., context representations). Model 1's assumptions are met, whereas GNB assumptions are violated. (Right) X is perfectly yoked to experiment condition, but R fluctuates according to a latent variable. This violates model 1's assumptions, but meets the GNB's assumptions.

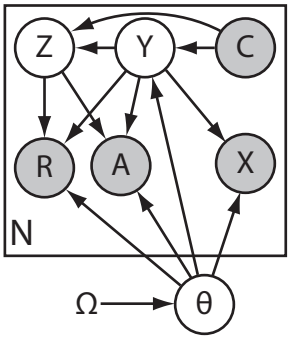



Figure 5.8: Generative model 2, following graphing convention in Bishop (2007). Variable definitions given in main text.

## 5.5 Model 2: Behavioral differences across cognitive states

Model 2 extends Model 1 by specifying direct links between cognitive state and behavior. These links are likely to exist in many datasets, and specifying them may therefore be important for reducing model misspecification in order to obtain more accurate parameter estimates and increase sensitivity. I derive the EM algorithm and cross-validation results in this section.

### 5.5.1 Generative Model Specification

Model 2 variables are identical to Model 1. The only differences between the models derive from the additional links in Model 2 between behavior variables (RT and accuracy) and cognitive state. Model 2 is depicted in Figure 5.8.

The likelihood function is given by:

$$\begin{aligned} p(\mathbf{U}, \mathbf{C}|\theta) &= \prod_{n=1}^{N} p(u_n, c_n|\theta) \\ &= \prod_{n=1}^{N} \sum_{y_n} p(x_n|y_n, \theta) p(y_n|c_n, \theta) \\ &\quad \cdot \sum_{z_n} p(r_n|y_n, z_n, \theta) p(a_n|y_n, z_n, \theta) p(z_n|y_n, c_n) p(c_n) \end{aligned}$$

The model parameters enter in as:

$$\begin{aligned} p(x|y, \theta) &= \mathcal{N}_D\left(x|\mu_y, \Sigma_y\right) \\ p(r|y, z, \theta) &= \text{Gamma}(r|\nu_y, \lambda_{yz}) \\ p(a|y, z, \theta) &= \text{Bernoulli}(\xi_{yz}) \\ p(y|c, \theta) &= \text{Multinomial}(y|\pi(c)), \quad \pi_i(c) = \begin{cases} \eta, & c_i = 1 \\ (1-\eta)\frac{1}{J-1}, & c_i = 0 \end{cases} \\ z|y, c &= \mathbb{I}\{y = c\} \\ p(c) &= 1/K \end{aligned}$$

The prior is given as:

$$\begin{aligned} p(\theta|\Omega) &= p(\eta|\Omega) \prod_{j=1}^{J} p(\mu_j, \sigma_j^2|\Omega) p(\nu_j|\Omega) \prod_{i=1}^{2} p(\lambda_{ij}|\Omega) p(\xi_{ij}|\Omega) \\ &= p(\eta|\eta_0) \prod_{j=1}^{J} \prod_{d=1}^{D} p(\mu_{jd}, \sigma_{jd}^2|\mu_0, \kappa_0, \alpha_0, \beta_0) p(\nu_j|\nu_0) \prod_{i=1}^{2} p(\lambda_{ij}|\tau_0, \lambda_0) p(\xi_{ij}|\xi_0) \\ &= \text{Beta}(\eta|\eta_0) \prod_{j=1}^{J} \prod_{d=1}^{D} \text{NG}(\mu_{jd}, \sigma_{jd}^2|\mu_0, \kappa_0, \alpha_0, \beta_0) \text{U}(\nu_j|0, \nu_0) \\ &\quad \cdot \prod_{i=1}^{2} \text{Gamma}(\lambda_{ij}|\tau_0, \lambda_0) \text{Beta}(\xi_{ij}|\xi_0) \end{aligned}$$

### 5.5.2 Parameter Estimation

#### E-Step

During the E-step, we update the probability distribution over latent variables, $q(\mathbf{Y}) = p(\mathbf{Y}|\mathbf{X}, \mathbf{R}, \mathbf{A}, \mathbf{C})$. As with the previous mixture models, this reduces to independent equations for each observation, as follows:

$$\begin{aligned} q(y_n) &= p(y_n|u_n, c_n, \theta) \\ &= Q^{-1} p(u_n|y_n, c_n, \theta) p(y_n|c_n, \theta) \\ &= Q^{-1} \sum_{z_n} p(u_n|y_n, z_n, c_n, \theta) p(z_n|y_n, c_n) p(y_n|c_n, \theta) \\ &= Q^{-1} p(x_n|y_n, \theta) \sum_{z_n} p(r_n|y_n, z_n, \theta) p(a_n|y_n, z_n, \theta) p(z_n|y_n, c_n) p(y_n|c_n, \theta), \\ Q &= p(u_n, c_n|\theta) \end{aligned}$$

The result is quite similar to model 1, except that RT and accuracy now depend on latent cognitive state as well as the match/mismatch between this and experiment condition.

#### M-Step

The M-step is essentially the same as for model 1, except for the behavior parameter estimates. These differ only in that we estimate parameters separately for each level of cognitive state. Assuming that common hyperparameters are used for all cognitive states, the accuracy parameters for the $j$-th cognitive state and the $i$-th value of $z$

are given by:

$$\hat{\xi}_{ij} = \frac{f_1 - 1}{f_1 + f_2 - 2}$$

$$f = \xi_{0i} + \left( \sum_{n=1}^{N} q(y_{nj}, z_n = i) a_n, \sum_{n=1}^{N} q(y_{nj}, z_n = i)(1 - a_n) \right)^T$$

$$q(y_{nj}, z_n = 1) = q(y_{nj}) c_{nj}, \quad q(y_{nj}, z_n = 0) = q(y_{nj})(1 - c_{nj})$$

The estimates for the Gamma RT density are analogous to those for model 1, with a separate estimate for each latent cognitive state. For the $j$-th cognitive state, we have:

$$\hat{\lambda}_{\nu j} = \left( \frac{f_1 - 1}{g_1}, \frac{f_2 - 1}{g_2} \right)^T$$

$$f = \tau_0 + \left( \sum_{n=1}^{N} q(y_{nj}, z_n = 1)\nu, \sum_{n=1}^{N} q(y_{nj}, z_n = 0)\nu \right)^T$$

$$g = \lambda_0 + \left( \sum_{n=1}^{N} q(y_{nj}, z_n = 1) r_n, \sum_{n=1}^{N} q(y_{nj}, z_n = 0) r_n \right)^T$$

$$\hat{\nu}_j = \arg\max_{\nu} \left\{ \sum_{n=1}^{N} q(y_{nj}) \log \sum_{z_n} p(r_n | z_n, \hat{\lambda}_\nu, \nu) p(z_n | y_{nj}, c_n) + \log p(\hat{\lambda}_\nu, \nu | \tau_0, \lambda_0, \nu_0) \right\}$$

Again, I constrain $\nu$ to be constant across values of $z$, although it is now free to differ across levels of the latent cognitive state, $y$.

**Log Posterior and Convergence**

The log posterior is given as:

$$\begin{aligned}\log p(\theta|\mathbf{U},\mathbf{C},\Omega) &= \log p(\mathbf{U}|\mathbf{C},\theta) + \log p(\theta|\Omega) + Q \\ &= \sum_{n=1}^{N} \log \sum_{y_n} p(x_n|y_n,\theta) \\ &\quad \cdot \sum_{z_n} p(z_n|y_n,c_n)p(r_n|y_n,z_n,\theta)p(a_n|y_n,z_n,\theta)p(y_n|c_n,\theta) \\ &\quad + \log p(\theta|\Omega) + Q\end{aligned}$$

As stated previously, any terms that are independent of $\theta$ are included in $Q$ and can be ignored when checking the log posterior for convergence.

### 5.5.3 Model Inversion and Cross-validation

Omitting dependence on estimated $\theta$ for clarity, the predictive distributions are:

$$\begin{aligned}p\left(c_n|u_n\right) &= Q_1^{-1} \sum_{y_n} p(x_n|y_n) \sum_{z_n} p(z_n|y_n,c_n)p(r_n|y_n,z_n)p(a_n|y_n,z_n)p(y_n|c_n)p(c_n) \\ Q_1 &= p(u_n) \\ \mathbb{E}[x_n|r_n,a_n,c_n] &= \sum_{y_n} \mu_{y_n} Q_2^{-1} \sum_{z_n} p(r_n|y_n,z_n)p(a_n|y_n,z_n)p(z_n|y_n,c_n)p(y_n|c_n) \\ Q_2 &= p(r_n,a_n|c_n) \\ \mathbb{E}[r_n|x_n,c_n] &= \sum_{y_n}\sum_{z_n} \frac{\nu_{y_n}}{\lambda_{y_n z_n}} p(z_n|y_n,c_n) Q_3^{-1} p(x_n|y_n)p(y_n|c_n) \\ Q_3 &= p(x_n|c_n)\end{aligned}$$

Test statistics for cross-validation are then constructed in the same way as with model 1. As with that model, it is again important to test each null hypothesis before interpreting results.

### 5.5.4 Model 2 Discussion

Model 2 generalized model 1 to include direct dependencies between cognitive state and behavior. Although this results in a slightly more complex model, it is intuitive that such dependencies exist, and this approach will reduce misspecification in that case while incurring only minor additional computational costs.

## 5.6 Additional Simulations

I described the GNB and several model-based MVPA variants above. Here I present additional simulations that compare the methods in various settings. The first set of simulations compares all the methods (the GNB, and models 0, 1, and 2) across different data generating scenarios. This demonstrates model properties in a side-by-side comparison, and demonstrates the need for thorough testing (multiple hypothesis rejection) in the more complex models. It also demonstrates performance with more realistic noise (see below). The second set of simulations compares all the methods as noise is degraded. This is useful, because model-based methods are nonlinear models that are fit by iterative methods, and thus may exhibit more rapid performance degradation as SNR decreases.

Which variables drove voxels and behavior — experiment condition, cognitive state, or null — was varied across simulations. The noise in the preliminary simulations presented above was uncorrelated across examples. To better reproduce the noise profile of an fMRI time series, the simulated voxel noise here additively combined white, Brownian, and auto-regressive components. Noise was uncorrelated across voxels, RT, and accuracy. Each simulated voxel time series was then high-pass filtered and mean centered in order to simulate fMRI data preprocessing. The simulated behavior data came from gamma (RT) and Bernoulli (accuracy) distributions.

Figure 5.9 shows results from the first simulation, which compared cross-validation

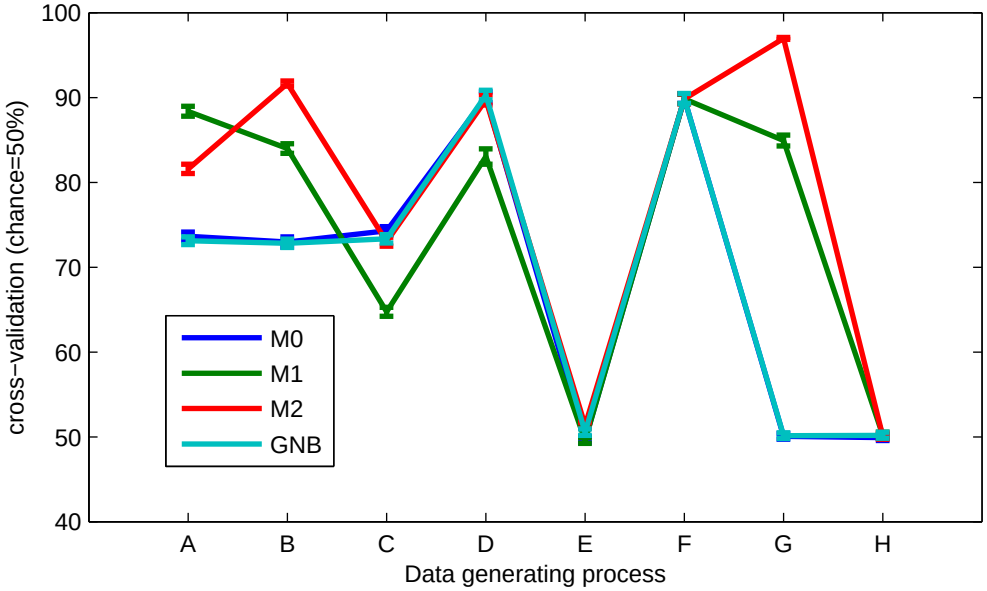



Figure 5.9: Cross-validation performance of different MVPA methods across different data generating processes. (A) Voxels (X) and behavior (R) both fluctuate relative to experiment condition (T), and fluctuate together according to a common latent variable as specified in model 1. This meets assumptions of model 1 and model 0. (B) X and R fluctuate relative to T, and both fluctuate according to a latent as specified by model 2. This meets the assumptions of model 2 and model 0. (C) X and R fluctuate relative to T, but each according to a unique latent variable. This meets assumptions of model 0. (D) X is reliably driven by T, but R fluctuates. This meets assumptions of model 0 (with zero probability of fluctuation) and the GNB. (E) X is just noise, R fluctuates. This does not meet any model assumptions. (F) X is reliably driven by T, R is just noise. This meets the assumptions of all models (with zero probability of fluctuation) except model 2. (G) X is just noise, R is reliably driven by T. This violates assumptions of all models. (H) Both X and R are just noise, violating assumptions of all models.

performance across methods and across different data generating processes. As expected, each method performed best when its generative model matched the true data generating process. The figure also shows that methods with more specific generative models were more selective, sacrificing general sensitivity while gaining sensitivity in the specific situations for which they were designed. In particular, the behavior-sensitive models 1 and 2 excelled relative to the behavior-insensitive GNB and model 0 in the cases where behavior and voxels met their assumptions. In contrast, behavior-sensitive methods performed somewhat poorly relative to the others when there was a behavioral signal that violated model assumptions. This demonstrates in a range of situations the principle noted above, namely that models with more assumptions become more specifically sensitive. Behavioral data that violated assumptions "distracted" these methods, leading to distorted parameter estimates. In contrast, the GNB and basic mixture methods assume nothing about behavior, and therefore their performance was unaffected by the structure of the behavioral data, yielding consistent, moderate performance whenever there was a voxel signal. One apparent exception to the rule of generative models changing sensitivity is the fact that model 0 performance is indistinguishable from GNB performance. However, as described above however, model 0 exhibits different performance than the GNB when using other generalization metrics, such as pattern separability. Finally, Figure 5.9H demonstrates the point that more complex models must be tested more thoroughly — despite the fact that the voxels contain only noise, both model 1 and model 2 predict experiment condition successfully during cross-validation. This is due to the fact that, in this particular simulated dataset, behavior is directly related to experiment condition. Both model 1 and model 2 link behavior to task in such a way that parameter estimation can be "hijacked" in this scenario, leading to successful experiment condition cross-validation. This is simple to check, in practice: predicting voxels during cross-validation naturally fails in this same scenario. Therefore, form-

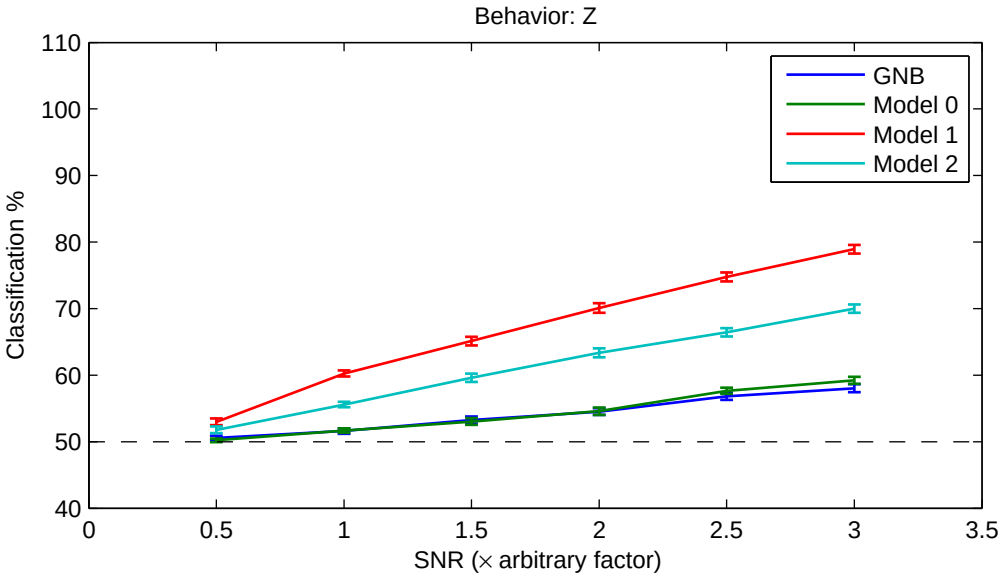



Figure 5.10: Experiment condition cross-validation performance of different methods across SNR levels. Data were generated according to model 1 assumptions: voxels are driven by cognitive state that fluctuates relative to experiment condition, and behavior is driven by match/mismatch between cognitive state and experiment condition.

ing a conjunction map of experiment condition and voxel cross-validation produces interpretable results.

The second simulation (Figure 5.10) explores the robustness of different methods across a range of voxel signal-to-noise ratio (SNR) levels. This was an important simulation, because the model-based methods are all mixture models with iterative parameter estimation procedures that are subject to local optima. Thus it is important to compare their robustness to noise to that of the GNB. However, this simulation showed that the model-based methods degraded gracefully with SNR, maintaining equal or slightly superior performance relative to the GNB as SNR decreased toward chance level. It is particularly important to note that this was true for Model 0, as this was the only model-based method that could not benefit from the behavioral signal in this simulation.

## 5.7 Application to fMRI Data

In the previous chapter, I argued that recently emerging reports (Bode & Haynes, 2009; Reverberi et al., 2011; Woolgar et al., 2011; Cole et al., 2011) are difficult to interpret due to the fact that context is systematically confounded with task difficulty under standard, classification-based MVPA mapping methods. I described newly collected data that replicated previous context representation findings with standard MVPA. However, controlling for the confounds by a fairly crude linear regression method eliminated these results. This was part of the motivation for developing the model-based MVPA methods described in the present chapter. In this section, I report on the application of the model-based MVPA methods to the fMRI data described previously, and I provide a comparison between model-based and standard MVPA methods.

### 5.7.1 Methods

All methods except for model-based MVPA analysis are as described in the previous chapter. For comparison purposes, a support vector machine (SVM) discriminative classifier was used. For the SVM analysis, all preprocessing was identical to GNB preprocessing. Following the last chapter, trial-to-trial RT variability was again removed from the voxel data by linear regression prior to analysis, *except* for the model-based methods. It is not necessary to remove this confound from the model-based methods, as the conjunction of multiple cross-validation maps and specificity of the model-based methods controls for it automatically.

### 5.7.2 Results

As with the previous chapter, context representation mapping produced null results across both experiments. No analysis revealed significant differences between task 1

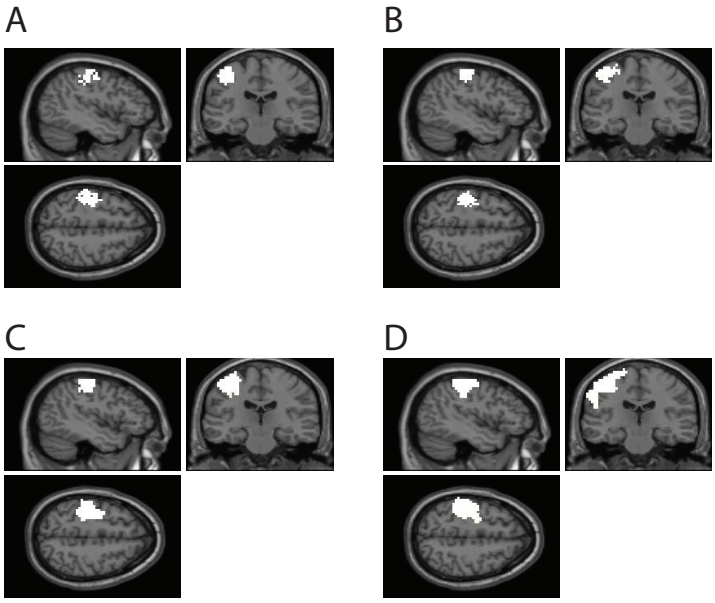



Figure 5.11: Experiment 1. Left/Right response analysis produces results in L post-central gyrus with all analysis methods. All views are centered on MNI coordinates -45x, -22y, 54z. All white voxels belong to local clusters that are significant at the 0.05 level after correction for wholebrain multiple comparisons. (A) SVM classifier identifies 204 significant voxels. (B) Mixture model classifier identifies 246 significant voxels. (C) GNB classifier identifies 493 significant voxels. (D) GLM contrast (right minus left contrast) identifies 672 significant voxels.

and task 2 activation anywhere in the brain. In contrast, for the left/right response analysis, GLM, GNB, SVM, and model 0 methods all located activation differences in left postcentral gyrus. This is shown for Experiment 1 in Figure 5.11. Model 0 was used in the left/right response analysis whereas models 1 and 2 were not. This was because model 0 makes few assumptions and was likely to produce results comparable with the GNB. Models 1 and 2 would produce little information when applied to this analysis, due to their relative specificity and the lack of *a priori* hypothesis that they would describe the left/right representations.

## 5.8 General Discussion

### 5.8.1 Cross-validation versus similarity analysis

I have presented models in this chapter with an emphasis on using prediction-based cross-validation such as experiment condition prediction for inferential tests. This is a well-established general statistical practice, and supports a broad treatment of the model-based MVPA methods developed here. It is also most closely related to classical, mass-univariate fMRI analysis techniques. Nonetheless, there is a newer approach to MVPA hypothesis testing that bears consideration here, namely RSA (e.g., Kriegeskorte et al., 2008b, 2008a). RSA takes as its basic unit the correlations among patterns of voxel activity in different experiment conditions. The structure among the between-pattern correlations can then be related to various cognitive hypotheses. For example, consider an experiment with four tasks (i.e., experiment conditions in this example) with varying degrees of overlap among stimulus-response mappings, so that tasks A and B share three stimulus-response mappings, tasks A and C share two stimulus-response mappings, and tasks A and D share one stimulus-response mapping. If the pattern across voxels is estimated for each task, then one could use the Fisher transformation to compute the following z-scored between-pattern correlations:

$z_{AB}$, $z_{AC}$, and $z_{AD}$. One could then test the hypothesis that correlations among these patterns are driven by overlap in stimulus response mappings by testing the contrast $(1, 0, -1)(z_{AB}, z_{AC}, z_{AD})^T$ against zero. Areas of the brain where groups of voxels reject this null hypotheses would be said to exhibit activity patterns consistent with the representation of stimulus-response mappings.

There are several interesting advantages of RSA relative to cross-validation analysis. For example, RSA does not necessarily require a held out test set. Instead, inferential tests may sometimes be conducted on the same data from which parameters were estimated. This can reduce total computation time by an order of magnitude by eliminating cross-validation folds. Moreover, with similarity analyses it is unnecessary to identify test statistics for each predicted variable that are symmetrically distributed about some known value. Rather, zero-centered Gaussian test statistics arise naturally from the null hypotheses associated with RSA.

Perhaps the most interesting advantage of similarity analyses with regard to the model-based MVPA methods developed in this chapter is that they permit direct testing of cognitive hypotheses about the voxel patterns themselves, rather than indirect testing by way of cross-validated prediction. As noted above, this is particularly interesting with regard to model 0. Figure 5.1E suggests that a model that captures fluctuations but does not include behavior, such as model 0, should be more successful than the GNB in recovering the parameter estimates, or the pattern across voxels, associated with cognitive states. Such a model would not necessarily be as successful as a model that correctly specified behavior in addition to fluctuations, but it would exhibit a broader sensitivity profile due to the lack of explicit assumptions regarding behavior. Simulations demonstrated above that this potential is not realized with cross-validation tests, but is realized with a pattern separability statistic, which is related to an RSA test statistic. Thus similarity analysis may provide an avenue for more thorough exploration of models such as model 0 that include separate cogni-

tive state variables but do not make assumptions about behavior. For example, this opens up the possibility of exploring the "$J > K$" space of models with more cognitive states than experimental conditions. For example, $J = K + 1$ could specify a model with a non-specific "mind wandering" state. Although this seems to be a plausible way to reduce model misspecification in principle, in practice it would exhibit poor sensitivity in prediction-based cross-validation inferential tests. In contrast, it might be very sensitive when used with RSA-like tests.

In sum, although I have emphasized standard cross-validation inferential tests in this chapter, RSA-like tests may reveal additional benefits of the model-based MVPA approach. In particular, similarity analysis may be well suited to realize the benefits for models that include distinct (i.e., latent) cognitive state variables, but do not include behavior. All of the model specification and parameter estimation results presented in this chapter would remain applicable for RSA. However, cross-validation tests would be replaced with other methods. The reader is referred to Kriegeskorte et al. (2008b, 2008a) for more information.

### 5.8.2 Connectionist Implementations of GAT

GAT is typically implemented with neural network or connectionist models (e.g., Cohen et al., 1990; Rougier et al., 2005), and this is a natural computational framework for implementing some of features of the theory. Yet when I set out to implement GAT in model-based MVPA, I quickly moved toward gross simplifications based on exponential family and mixture-of-exponential-family statistical models. This led to model-based MVPA classifiers that were "inspired" by GAT, capturing some of the core principles but few of the details. As described above, the motivation for this gross simplification is that the MVPA application places rigorous demands on candidate generative models. These are: (1) the model must include voxels as well as behavior; (2) there must be an efficient procedure for estimating model parameters; and (3) the

model must be efficiently "invertible" in order for cross-validation to work, meaning that, alone and together, voxels and behavior must be informative about the latent cognitive state, which in turn must be informative about about experiment condition.

It is difficult to meet these demands with a connectionist implementation of GAT. Whereas MVPA requires that a generative model include individual voxels, connectionist GAT models typically include behavior alone, or else include behavior together with low-dimensional statistics of neural data (e.g., the mean activation in a region) rather than individual voxel data. Individual voxel components could be included, and would presumably be integrated into the network's task layer, at the expense of including one or two additional model parameters (i.e., mean, or mean and variance) per voxel.

Parameter estimation is a perhaps the greatest challenge facing the use of connectionist models in model-based MVPA. As generative models of voxels and behavior, connectionist GAT models are not mixtures of exponential family models (unlike all models presented above) that would have closed form MAP or MLE solutions. Therefore, the EM algorithm cannot be used effectively, and a general purpose gradient ascent algorithm would be necessary. Moreover, it is extremely difficult to approximate the likelihood function of a connectionist GAT model. Indeed, these models do not explicitly express the likelihood function at all, but rather provide a mechanism for drawing random samples from it. Thus MAP and MLE are likely to be impractical with these models, meaning that parameters would usually be estimated by the method of moments (i.e., by finding parameters that minimize the distance between the model's simulated moments and the data's moments). For a given parameter point, this proceeds by simulating many samples from the implicit likelihood (the joint over observable variables) and then averaging these samples to form Monte Carlo approximations of the model's moments, $\mathbb{E}[U|\theta], \mathbb{E}[U^2|\theta], \mathbb{E}[U^3|\theta], ..., \mathbb{E}[U^L|\theta]$, where $L$ is the number of model parameters. Parameters are then adjusted to de-

scend the gradient of the distance metric (e.g., sum of squared errors) between the simulated moments and the data's moments. This process is computationally expensive. Each sample from the likelihood is itself computationally expensive, and many samples must be drawn in order to simulate the model's moments with sufficient precision for gradients to be computed. Moments must be simulated at many points in a neighborhood of parameter space in order to estimate local first and second derivatives. The number of local points necessary here increases quickly with the number of parameters. The best way to deal with this is to supply the first and second derivatives of the cost function. However, with connectionist GAT models this cost function is the distance metric between model and data moments, which is a stochastic function that is non-differentiable. To clarify, if each sample of the likelihood was deterministic and only the final simulated moment was stochastic, then the first and second derivatives could be computed for each sample and effectively combined at the end of the simulation. However, each sample is itself stochastic, as this is an essential part of the GAT model of, e.g., RT and accuracy (e.g. Liu, Holmes, & Cohen, 2008).

As demonstrated in simulations above, gradient methods scale poorly with the number of parameters, even when supplied with the first derivative. Parameter scalability is an important problem with model-based MVPA, because the need to model individual voxels often implies a large number of parameters. In sum, parameter estimation is difficult with connectionist GAT models even when they include only behavior (and thus on the order of just 10 parameters). Adding the hundreds of parameters required to model voxel data would increase the computation time required to draw each sample, the number of samples required to precisely simulate the central moments, and the number of parameter points visited by the gradient ascent algorithm for each seed point. To build intuition about specific numbers, recall that in Figure 5.4, a gradient ascent method required over 450 seconds to estimate parameters for a model with 256 voxels, when the cost function (the model's log likelihood

function) took approximately one millisecond to compute. In contrast, the distance between model and data moments in a connectionist GAT model takes on the order of one second to compute on the same hardware, so for the same problem, the gradient ascent algorithm would probably take more than 450,000 seconds to converge, since in the connectionist case no analytic gradient could be supplied.

Another challenge is that some critical model inversions are difficult to achieve with connectionist GAT models. The reason is that these models are designed to produce samples from the likelihood, $p(r_n, a_n|c_n, \theta)$ by fixing the experiment condition variable and then simulating the neural network's dynamics to sample from the likelihood over $r_n$ and $a_n$. This is not a symmetric process: it is not obvious how to clamp RT and accuracy. That is, the model does not variables labeled "RT" or "accuracy" in the same sense that it has a variable labeled "experiment condition." Experiment condition is represented by actual nodes in the neural network, whereas the behavioral variables arise as statistics of a sample of the network's stochastic dynamics.

Although there are aspects of GAT that are specific to connectionist implementations, there are several challenges that must be overcome in order to use such implementations as generative models in MVPA classifiers. One promising approach would be to employ a two-step process in which a connectionist GAT model was first fit to behavior alone, and was then queried to produce the distribution over context representations on each trial, $p(y_n|r_n, a_n, c_n, \theta)$. This distribution could then be included as a covariate in standard (not model-based) MVPA. In other words, this approach would first use a connectionist GAT model to extract latent cognitive state information from behavior, and would then use this extracted information in a simple linear model of voxels. This would be particularly relevant for "searchlight" type analyses, as it would circumvent the challenge of needing to fit the connectionist model thousands of times, once for each local voxel neighborhood. With regard

to the goal of literally using connectionist models in model-based MVPA, the most important step is probably to reduce the dimensionality of the voxel data during preprocessing. This would avoid the immense computational cost of fitting a connectionist neural network model with hundreds or thousands of parameters. This was not considered in this chapter, as it represents an additional level of departure from standard MVPA approaches that would need to be explored and justified. However, if this approach is justified in future work, it may be readily achievable using standard, efficient dimensionality reduction methods such as principal components analysis.

### 5.8.3 Application to fMRI Data

I described the application of model-based MVPA methods to the context representation datasets initially described in the previous chapter. Unfortunately, none of the model-based MVPA methods here improved on the null results produced by basic methods. Like other methods, model-based MVPA methods detected neural differences between left and right behavioral responses in left postcentral gyrus.

It is always difficult to interpret null results, and the present failure to detect context representations may have been due to, for example, low signal-to-noise ratio. While it is disappointing that the specific model-based MVPA classifiers developed above did not produce significant context representation maps, this does not invalidate the theoretical work done here, or suggest that model-based MVPA is a poor approach to mapping context representations. Rather, it may be that the particular models developed above are too gross in their simplifying assumptions, and that further work developing on variations may reveal the usefulness of the theme. Specifically, I was at pains to stress that I was not arbitrary in my choice of simple generative models. Nonetheless, methods such as dimensionality reduction, and frankly, increasing computer speed, may mean that more complex models may soon be used in the spirit introduced here. It may well be that model-based MVPA classifiers with sub-

stantially richer generative models would be able to identify context representations in the present dataset.

### 5.8.4 Conclusions

Despite successes in the visual domain (e.g., Haynes & Rees, 2005; Kamitani & Tong, 2005; Freeman et al., 2011), there has been very little successful application of MVPA to identifying context representations in fMRI data. I have presented the a new family of model-based MVPA methods that may facilitate this goal. Model-based MVPA increases specificity and sensitivity relative to standard MVPA methods by incorporating aspects of GAT into standard MVPA tools. The overall model-based MVPA approach is also general in that it can be used to incorporate other domain-specific cognitive models rather than GAT. I have described the implementation of several model-based MVPA classifiers, and have provided simulations showing that the approach does indeed increase statistical sensitivity when the assumptions of the incorporated model are met by the data. I have also discussed shortcomings of this approach as well as possible avenues for addressing these in future work. In the next chapter, I will describe the application of this and other approaches to actual fMRI data.

# Chapter 6

# Discussion

Task representations, or context-sensitive stimulus-response mappings, form the central, necessary support for flexible, goal-directed behavior. As discussed in the introduction, GAT (Cohen et al., 1990; Cohen & Servan-Schreiber, 1992; Miller & Cohen, 2001) is a family of models that describes task representations as emerging from interactions between context representations in PFC and stimulus-response mapping representations in posterior cortical and subcortical structures. Specifically, GAT assumes that stimulus-response mappings are excited in a context-insensitive manner by bottom-up stimulus representations and compete for control of behavior, but context representations provide top-down support for the goal-relevant mapping by biasing the competition with long-range, excitatory projections. Gating models extend GAT to address the stability-flexibility paradox of context representations themselves. These models posit that RL processes — mediated by phasic dopamine bursts (Braver & Cohen, 1999, 2000) and corticostriatal loops (Frank et al., 2001; O'Reilly & Frank, 2006; Todd et al., 2008) — derive useful context updating (i.e., gating) signals from primordial reward and punishment signals. Taking GAT and the Gating models together, it is assumed that task representations emerge from context representations in PFC that select stimulus-response mappings in posterior cortical

and subcortical structures, and are in turn controlled by dopaminergic and basal ganglia RL processes. In the preceding chapters, I described several projects that sought to extend and test these models.

I first described a project that followed up on a recent Gating model (O'Reilly & Frank, 2006) by establishing new links to RL theory. This allowed for the biologically inspired but cumbersome model to be simplified into a three-parameter form, and also provided an RL-based theoretical definition of working memory (of which task representations are one form) as the "belief-state" variable required by agents to navigate partially observable environments. The emphasis on working memory representations as a means to learning optimal behavior in complex environments also led to suggestions for why working memory capacity is so limited (e.g. Cowan, 2000), namely that the curse of dimensionality (i.e., geometric increase in learning time as a function of the capacity of working memory) sharply limits the number of items that can be usefully stored.

The second project, I explored the validity of a proposed normative account of task representation updating at the within-trial timescale in the task switching domain. Specifically, De Jong (2000) and Nieuwenhuis and Monsell (2002) previously proposed that the timing of task representation updating (as measured by the so-called residual task switch cost) is well described by an effort-performance tradeoff, with earlier updating corresponding to greater effort but improved performance, and later updating corresponding to lower effort and worse performance. However, there was a speed-accuracy tradeoff in at least some of data used to support that proposal, and this was not accounted for in the original arguments. This introduced the possibility that the observed effects arose not from an effort-performance tradeoff, but merely from a speed-accuracy tradeoff, undermining support for the proposed normative account. Since the potential for complex interactions between speed-accuracy tradeoffs and effort-performance tradeoffs clearly existed, I first formalized the effort-performance

tradeoff and competing speed-accuracy tradeoff proposals within the rational analysis framework (Anderson, 1990, 1991). This afforded precise articulation of the competing hypotheses. I then established that both hypotheses could account for existing data, and I conducted new experiments to better test between them. Both experiments provided support for the model that included a speed-accuracy tradeoff alone, without an effort-performance tradeoff. I concluded that the proposed normative account of task representation updating receives no support from the data, and should be revised.

In a third project, I attempted to map context representations in fMRI data. In the first part of this project, I conducted two new fMRI experiments in order to search for context representations (hypothesized to reside in PFC), following up on similar recent work (Bode & Haynes, 2009; Cole et al., 2011; Reverberi et al., 2011; Woolgar et al., 2011). My work revealed systematic confounds inherent in the standard MVPA methods used in the recent previous studies. I showed that, although the new experimental data initially agreed with the recent results, controlling for these confounds eliminated the results. Control analyses (i.e., differentiating between left and right motor responses) succeeded after controlling for confounds, demonstrating the basic viability of the datasets. The recognition of confounds and the ensuing null context representation results motivated the development of novel, model-based MVPA methods. The goal of model-based MVPA is to increase sensitivity and specificity (i.e., decrease sensitivity to confounds) by incorporating aspects of GAT, including links between trial-to-trial behavioral variability and context representations, into the standard GNB MVPA classifier. Theory and simulations showed that this approach does indeed increase sensitivity in brain regions where the voxels interacted with behavior and experiment condition according to the model, while decreasing sensitivity in brain areas where some but not all of the model's assumptions are met. Ultimately, the model-based MVPA methods were used to re-analyze the dataset, and did not

produce new results. However, the model-based MVPA approach appears to be fairly general and theoretically powerful, and future developments in that area may yet lead to sufficient sensitivity for detecting context representations.

In conclusion, GAT and the Gating models are prominent descriptions of task representations, and I have presented several projects that attempted to further these lines of work. In one project, I further developed the theory underlying the Gating models. In another project, I showed that a proposed normative account of task representation updating is unsupported by the data and needs to be revised. In a third project, I attempted to map context representations (part of task representations) using fMRI. I explored methodological limitations with previous work, and developed new model-based methods that showed theoretical promise in addressing these limitations.

# References


Allport, A., Styles, E., & Hsieh, S. (1994). Shifting intentional set: exploring the dynamic control of tasks. *Attention and performance XV*, *15*, 421–452.

Altmann, E., & Gray, W. (2008). An integrated model of cognitive control in task switching. *Psychological Review*, *115*(3), 38.

Anderson, J. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Badre, D., & Wagner, A. D. (2004). Selection, integration, and conflict monitoring; assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron*, *41*(3), 473–87.

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*(2), 315–26. doi:10.1016/j.neuron.2010.03.025

Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(04), 723–742.

Barch, D. M., Braver, T. S., Nystrom, L. E., Forman, S. D., Noll, D. C., & Cohen, J. D. (1997). Dissociating working memory from task difficulty in human prefrontal cortex. *Neuropsychologia*, *35*(10), 1373–80.

Bertsekas, D., & Tsitsiklis, J. (1996). *Neuro-dynamic programming*. Athena Scientific (Belmont, Mass.)

Bettman, J., Johnson, E., & Payne, J. (1990). A componential analysis of cognitive effort in choice. *Organizational behavior and human decision processes(Print)*, *45*(1), 111–139.

Bishop, C. (2007). *Pattern recognition and machine learning (information science and statistics)*. Springer.

Bode, S., & Haynes, J.-D. (2009). Decoding sequential stages of task preparation in the human brain. *Neuroimage*, *45*(2), 606–13. doi:10.1016/j.neuroimage.2008.11.031

Bogacz, R., Brown, E. T., Moehlis, J., Holmes, P. J., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev*, *113*(4), 700–65. doi:10.1037/0033-295X.113.4.700

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol Rev*, *108*(3), 624–52.

Botvinick, M. M., Wang, J, Cowan, E, Roy, S, Bastianen, C, Mayo, J. P., & Houk, J. (2009). An analysis of immediate serial recall performance in a macaque. *Anim Cogn*. doi:10.1007/s10071-009-0226-z

Boynton, G. M. (2005). Attention and visual perception. *Curr Opin Neurobiol*, *15*(4), 465–9. doi:10.1016/j.conb.2005.06.009

Brainard, D. (1997). The psychophysics toolbox. *Spatial vision*, *10*(4), 433–436.

Braver, T. S., & Cohen, J. D. (1999). Dopamine, cognitive control, and schizophrenia: the gating model. *Prog Brain Res*, *121*, 327–49.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: the role of dopamine in regulating prefrontal function and working memory. *Attention and performance XVIII*, 713–737.

Brown, J. W., Reynolds, J. R., & Braver, T. S. (2007). A computational model of fractionated conflict-control mechanisms in task-switching. *Cogn Psychol*, *55*(1), 37–85. doi:10.1016/j.cogpsych.2006.09.005

Brown, S., Lehmann, C., & Poboka, D. (2006). A critical test of the failure-to-engage theory of task switching. *Psychonomic bulletin & review*, *13*(1), 152–159.

Bunge, S. A. (2004). How we use rules to select actions: a review of evidence from cognitive neuroscience. *Cogn Affect Behav Neurosci*, *4*(4), 564–79. Retrieved from http://cabn.psychonomic-journals.org/content/4/4/564.long

Carlin, J., Calder, A., Kriegeskorte, N., Nili, H., & Rowe, J. (2011). A head view-invariant representation of gaze direction in anterior superior temporal sulcus. *Current Biology*.

Chiu, Y., Esterman, M., Gmeindl, L., & Yantis, S. (2011). Tracking cognitive fluctuations with multivoxel pattern time course (mvptc) analysis. *Neuropsychologia*.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *J Exp Psychol Gen*, *120*(3), 235–53. Retrieved from http://psycnet.apa.org/index.cfm?fa=search.displayRecord&uid=1992-00389-001

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychol Rev*, *99*(1), 45–77.

Cohen, J. D., Dunbar, K. O., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychol Rev*, *97*(3), 332–61.

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, *386*(6625), 604–8. doi:10.1038/386604a0

Cole, M., Etzel, J., Zacks, J., Schneider, W., & Braver, T. (2011). Rapid transfer of abstract rules to novel contexts in human lateral prefrontal cortex. *Frontiers in Human Neuroscience*, *25*(3), 607–23.

Cowan, N. (2000). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav Brain Sci*, *24*(1), 87–114; discussion 114–85.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation*, *18*(7), 1637–1677.

Dayan, P. (2007). Bilinearity, rules, and prefrontal cortex. *Front Comput Neurosci*, *1*(Article 1), 1–14. doi:10.3389/neuro.10.001.2007

Dayan, P., & Abbott, L. F. (2001). Theoretical neuroscience: computational and mathematical modeling of neural systems - chpt. 9: reinforcement learning. *MIT Press*. Retrieved from http://www.mitpressjournals.org/doi/pdf/10.1162/089892903321107891

De Jong, R. (2000). An intention-activation account of residual switch costs. In S. Monsell & J. Driver (Eds.), *Attention and performance xviii: control and cognitive processes* (pp. 357–376). MIT Press.

De Jong, R., Berendsen, E., & Cools, R. (1999). Goal neglect and inhibitory limitations: dissociable causes of interference effects in conflict situations. *Acta Psychologica*, *101*(2-3), 379–394.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu Rev Neurosci*, *18*, 193–222. doi:10.1146/annurev.ne.18.030195.001205

D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S, & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature*, *378*(6554), 279–81. doi:10.1038/378279a0

Dreisbach, G., & Haider, H. (2006). Preparatory adjustment of cognitive control in the task switching paradigm. *Psychon Bull Rev*, *13*(2), 334–8. Retrieved from http://www.ingentaconnect.com/content/psocpubs/pbr/2006/00000013/00000002/art00023?token=005f1f4ed8978a0be001fca644a467b4d616d3f4e4b34496e586546244044423b47665d36786c2545a271e5ab77b82a

Dreisbach, G., Haider, H., & Kluwe, R. H. (2002). Preparatory processes in the task-switching paradigm: evidence from the use of probability cues. *J Exp Psychol Learn Mem Cogn*, *28*(3), 468–83.

Droll, J., & Hayhoe, M. (2007). Trade-offs between gaze and working memory use. *Journal of Experimental Psychology*, *33*(6), 1352–1365.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*(11), 820–829.

Forman, S. D., Cohen, J. D., Fitzgerald, M, Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fmri): use of a cluster-size threshold. *Magn Reson Med*, *33*(5), 636–47.

Frank, M. J., Loughry, B, & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cogn Affect Behav Neurosci*, *1*(2), 137–60. Retrieved from http://www.ingentaconnect.com/content/psocpubs/cabn/2001/00000001/00000002/art00002?token=004f16b7855a373f6a4b6e4e395e4e6b63314f39412f415d482834447b2e4a6c7a403842449d2d3

Freeman, J., Brouwer, G., Heeger, D., & Merriam, E. (2011). Orientation decoding depends on maps, not columns. *The Journal of Neuroscience*, *31*(13), 4792.

Gauvain, J., & Lee, C. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *Speech and Audio Processing, IEEE Transactions on*, *2*(2), 291–298.

Gevins, A., Smith, M., McEvoy, L., & Yu, D. (1997). High-resolution eeg mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cerebral Cortex*, *7*(4), 374–385.

Gold, J. I. (2002). Good vibrations. *Neuron*, *33*(6), 842–4. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0896627302006293`

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A, Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex - supplemental material. *Science*, *293*(5539), 2425–30. Notes supplemental material (online). doi:`10.1126/science.1063736`

Haynes, J.-D., & Rees, G. (2005). Predicting the stream of consciousness from activity in human visual cortex. *Curr Biol*, *15*(14), 1301–7. doi:`10.1016/j.cub.2005.06.026`

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat Rev Neurosci*, *7*(7), 523–34. doi:`10.1038/nrn1931`

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Curr Biol*, *17*(4), 323–8. doi:`10.1016/j.cub.2006.11.072`

Houk, J., Adams, J., & Barto, A. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). The MIT Press, Cambridge, MA.

Jersild, A. (1927). Mental set and shift. *Archives of Psychology*, *89*.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, *15*(4-6), 535–547.

Johnson, E., & Payne, J. (1985). Effort and accuracy in choice. *Management science*, *31*(4), 395–414.

Johnson, J. D., McDuff, S. G. R., Rugg, M. D., & Norman, K. A. (2009). Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. *Neuron*, *63*(5), 697–708. doi:10.1016/j.neuron.2009.08.011

Kaelbling, L., Littman, M., & Cassandra, A. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1-2), 99–134.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nat Neurosci*, *8*(5), 679–85. doi:10.1038/nn1444

Kool, W., McGuire, J., Rosen, Z., & Botvinick, M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665.

Kriegeskorte, N., Goebel, R., & Bandettini, P. A. (2006). Information-based functional brain mapping. *Proc Natl Acad Sci USA*, *103*(10), 3863–8. doi:10.1073/pnas.0600244103

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc Natl Acad Sci USA*, *104*(51), 20600–5. doi:10.1073/pnas.0705654104

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008a). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–41. doi:10.1016/j.neuron.2008.10.043

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008b). Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci*, *2*(Article 4), 1–28. doi:10.3389/neuro.06.004.2008

Liu, Y. S., Holmes, P. J., & Cohen, J. D. (2008). A neural network model of the eriksen task: reduction, analysis, and data fitting. *Neural Comput*, *20*(2), 345–73. doi:10.1162/neco.2007.08-06-313

Loch, J., & Singh, S. (1998). Using eligibility traces to find the best memoryless policy in partially observable markov decision processes. In *Proceedings of the fifteenth international conference on machine learning* (pp. 323–331). Citeseer.

Logan, G. D., & Bundesen, C. (2003). Clever homunculus: is there an endogenous act of control in the explicit task-cuing procedure? *J Exp Psychol Hum Percept Perform*, *29*(3), 575–99. Retrieved from `http://psycnet.apa.org/index.cfm?fa=search.displayRecord&uid=2003-05350-006`

McCallum, R. (1995). Instance-based utile distinctions for reinforcement learning with hidden state. In *Machine learning-international workshop then conference-* (pp. 387–395). Citeseer.

McDuff, S. G. R., Frankel, H. C., & Norman, K. A. (2009). Multivoxel pattern analysis reveals increased memory targeting and reduced use of retrieved details during single-agenda source monitoring. *J Neurosci*, *29*(2), 508–16. doi:`10.1523/JNEUROSCI.3587-08.2009`

McGuire, J. T., & Botvinick, M. M. (2010). Prefrontal cortex, cognitive control, and the registration of decision costs. *Proc Natl Acad Sci USA*, *107*(17), 7922–6. doi:`10.1073/pnas.0910662107`

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1423–1442.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci*, *24*, 167–202. doi:`10.1146/annurev.neuro.24.1.167`

Monsell, S. (2003). Task switching. *Trends Cogn Sci*, *7*(3), 134–140.

Montague, P. R., Dayan, P, & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J Neurosci*, *16*(5), 1936–47.

Montojo, C. A., & Courtney, S. M. (2008). Differential neural activation for updating rule versus stimulus information in working memory. *Neuron*, *59*(1), 173–82. doi:10.1016/j.neuron.2008.05.012

Nieuwenhuis, S., & Monsell, S. (2002). Residual costs in task switching: testing the failure-to-engage hypothesis. *Psychon Bull Rev*, *9*(1), 86–92.

Nieuwenhuis, S., Ridderinkhof, K. R., Talsma, D., Coles, M. G. H., Holroyd, C. B., Kok, A., & van der Molen, M. W. (2002). A computational account of altered error processing in older age: dopamine and the error-related negativity. *Cogn Affect Behav Neurosci*, *2*(1), 19–36.

Norman, D., & Shallice, T. (1986). Attention to action: willed and automatic control of behaviour. In R. J. Davidson, G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation: advances in research and theory* (Vol. 4, pp. 1–18). New York: Plenum Press.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends Cogn Sci*, *10*(9), 424–30. doi:10.1016/j.tics.2006.07.005

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput*, *18*(2), 283–328. doi:10.1162/089976606775093909

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *J Neurosci*, *30*(30), 10127–10134. doi:10.1523/JNEUROSCI.2161-10.2010

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, *45*(1), S199–S209.

Peshkin, L., Meuleau, N., & Kaelbling, L. (1999). Learning policies with external memory. In *Sixteenth international conference on machine learning* (pp. 307–314).

Petrides, M. (2005). Lateral prefrontal cortex: architectonic and functional organization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 781.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*(5756), 1963–6. doi:10.1126/science.1117645

Puterman, M. (1994). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, Inc.

Quamme, J., Weiss, D., & Norman, K. (2010). Listening for recollection: a multi-voxel pattern analysis of recognition memory retrieval strategies. *Frontiers in Human Neuroscience*, *4*.

Rainer, G., Rao, S. C., & Miller, E. K. (1999). Prospective coding for objects in primate prefrontal cortex. *J Neurosci*, *19*(13), 5493–505.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Reverberi, C., Görgen, K., & Haynes, J. (2011). Compositionality of rule representations in human prefrontal cortex. *Cerebral Cortex*, Advance online publication. doi:10.1093/cercor/bhr200

Reynolds, J., & Wickens, J. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, *15*(4-6), 507–521.

Reynolds, J., Hyland, B., & Wickens, J. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*(6851), 67–70.

Rogers, R., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology-General*, *124*(2), 207–230.

Romo, R, Brody, C. D., Hernández, A, & Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, *399*(6735), 470–3. doi:10.1038/20939

Rougier, N., Noelle, D., Braver, T., Cohen, J., & O'Reilly, R. (2005). Prefrontal cortex and flexible cognitive control: rules without symbols. *Proceedings of the National Academy of Sciences*, *102*(20), 7338–7343.

Samejima, K., & Doya, K. (2007). Multiple representations of belief states and action values in corticobasal ganglia loops. *Annals of the New York Academy of Sciences*, *1104*(1), 213–228.

Savine, A. C., & Braver, T. S. (2010). Motivated cognitive control: reward incentives modulate preparatory neural activity during task-switching. *J Neurosci*, *30*(31), 10294–305. doi:10.1523/JNEUROSCI.2052-10.2010

Schneider, W., & Shiffrin, R. (1977). Controlled and automatic human information processing: 1. detection, search, and attention. *Psychological Review*, *84*(1), 1–66.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–9.

Shah, A., & Oppenheimer, D. (2008). Heuristics made easy: an effort-reduction framework. *Psychological Bulletin*, *134*(2), 207.

Singh, S., Jaakkola, T., & Jordan, M. (1994). Learning without state-estimation in partially observable markovian decision processes. In *Proceedings of the eleventh international conference on machine learning* (pp. 284–292). Morgan Kaufmann.

Smith, P. (1990). A note on the distribution of response times for a random walk with gaussian increments. *Journal of Mathematical Psychology*, *34*(4), 445–459.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*(1), 83–98. doi:10.1016/j.neuroimage.2008.03.061

Spector, A., & Biederman, I. (1976). Mental set and mental shift revisited. *The American Journal of Psychology*, *89*, 669–679.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. The MIT Press.

Todd, M., Niv, Y., & Cohen, J. (2008). Learning to use working memory in partially observable environments through dopaminergic reinforcement. In D. Koller, Y. Bengio, D. Schuurmans, L. Bottou & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 21).

Verbruggen, F., Liefooghe, B., Vandierendonck, A., & Demanet, J. (2007). Short cue presentations encourage advance task preparation: a recipe to diminish the residual switch cost. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 342–356.

Vickery, T., Chun, M., & Lee, D. (2011). Ubiquity and specificity of reinforcement signals throughout the human brain. *Neuron*, *72*(1), 166–177.

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, *411*(6840), 953–6. doi:10.1038/35082081

Wilcox, N. (1993). Lottery choice: incentives, complexity and decision time. *The Economic Journal*, *103*(421), 1397–1417.

Wilcox, N. (1994). On a lottery pricing anomaly: time tells the tale. *Journal of Risk and Uncertainty*, *7*(3), 311–324.

Williams, R. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, *8*(3), 229–256.

Woolgar, A., Thompson, R., Bor, D., & Duncan, J. (2011). Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. *NeuroImage*, *56*(2), 744–752.

Yeung, N., & Monsell, S. (2003). Switching between tasks of unequal familiarity: the role of stimulus-attribute and response-set selection. *J Exp Psychol Hum Percept Perform*, *29*(2), 455–69.

Yoshida, W., & Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. *Neuron*, *50*(5), 781–9. doi:10.1016/j.neuron.2006.05.006