Some rigorous results on the neural coding problem

by

Liam Paninski

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Center for Neural Science
New York University
May 2003

Eero P. Simoncelli

To sweet r

ACKNOWLEDGMENTS

I want to acknowledge the members of the Simoncelli lab: Eero, for his generosity and all his help; Odelia Schwartz, for many conversations on the LN model and her help on thesis formatting; Jenny Li, for her instrumental role in running the statistical methods seminar; John Pillow, for introducing me to the relaxing benefits of thinking about asymptotics; Tim Saint, for souvlaki; Rob Dotson, for fixing my computer problems when he could and yelling at me in an endearing way when he couldn't; Nicolas Bonnier, for adding a certain je ne sais quois to the lab environment; and Jose Acosta, for his introduction to the strange, beautiful world of Powerman 5000.

I am grateful to all the teachers I've had over the years, especially John Donoghue, Nicho Hatsopoulos, Stu Geman, and David Mumford at Brown, who each had an immense influence on my scientific worldview.

I thank the members of my various advisory committees over the last few years for their patience, especially Dan Tranchina, Larry Maloney, John Rinzel, Jonathan Victor, and Raghu Varadhan, for generously agreeing to read this work. Michael Hawken deserves special thanks for his support during my uncertain first year at CNS.

My collaborators, especially Matt Fellows and Brian Lau, for countless hours of entertainment. Matt and Brian deserve special credit for their roles in my discovery of mus modestus and the neutral milk phenomenon. Also

Monica Nagy for her great generosity in the collaboration cited above.

My family for all their love and support. Also the Sussmans and Buckleys for their generosity (including but not limited to many delicious meals and days — weeks, actually — of warm hospitality).

My friends at CNS, especially Stu Greenstein and Maeve O'Connell. I should also thank Lana Rosis for some very nice birthday presents.

I was generously funded by a predoctoral fellowship from the Howard Hughes Medical Institute during my time here.

Finally, Rachel, for everything.

TABLE OF CONTENTS

LIST OF FIGURES

Introduction

The search for efficient, accurate representations of the neural code has been a central problem in neuroscience for the better part of the past century [Adrian, 1926, Perkel et al., 1967, Rieke et al., 1997]. The basic question is simply stated: given some experimentally observable signal x (some sensory stimulus, or a certain type of movement) what is the probability of a given response y (say, that a given neuron will emit an action potential)? To put it more precisely, we want to estimate the conditional probabilities $p(y|x)$, for as large a set of observable signals x as possible. Stated this way, it is clear that the neural coding problem is fundamentally statistical: how do we estimate $p(y|x)$ given finite data?

The relevance of statistical methods for neuroscience has been demonstrated eloquently elsewhere. To list just a few classical examples: the elucidation of "hyperacuity" phenomena [Simmons, 1979, Shapley and Victor, 1986, Wandell, 1995], and the demonstration that the behavior of various sensory systems is limited by physical noise [Hecht et al., 1942, Bialek, 1987]; the application of Bayesian methods for

modeling otherwise apparently contradictory results in perceptual psychology [Knill and Richards, 1996, Weiss et al., 2002]; more recently, the application of well-defined probabilistic methods for the design of efficient neural prosthetic devices [Loizou, 1998, Shoham et al., 2003]; all of this work depended critically on simple but powerful statistical ideas. To take a somewhat wider view, if we try to solve problems like the ones that brains solve without using brains per se — computer vision, speech recognition, and machine learning are a few fields which immediately come to mind — statistical techniques become dominant.

Nevertheless, there remains a great deal of room for work on the statistical theory relevant for the neural coding problem. The basic problem is easily stated: we want to know $p(y|x)$ for all x, and there are too many possible x. Thus, to make progress, we have to do one of three things. First, we can relax our requirements: instead of estimating $p(y|x)$, which is too difficult, perhaps we can get away with estimating a few important functions of $p(y|x)$, and make inferences based on these functions alone. Second, along the same lines, we can try to fit some parametric model to $p(y|x)$; again, this reduces the intractably high-dimensional (what statisticians would call "nonparametric") original problem into a hopefully more manageable, low-dimensional problem. Finally, we can take a different tack and try to adaptively optimize the design of our experiment — that is, to only probe the system at those points x for which the associated responses y will specify the form of $p(y|x)$ in as precise a manner as possible.

We address specific instances of each of these three approaches here. First, we look at a class of methods for estimating mutual information, one of the most important functions of $p(x, y)$. Our second problem involves estimating a simple but powerful "cascade" model for stimulus-dependent neural activity given some high-dimensional signal x. Finally, we develop some of the necessary theory underlying a version of the adaptive experimental design idea outlined above.

Part 1: Estimation of entropy and mutual information

We present some new results on the nonparametric estimation of entropy and mutual information. The setup places no assumptions on the underlying probability measure generating the data; this generality is necessary for neural data, whose complexity and diversity complicates efforts to model these probability distributions parametrically with sufficient precision for accurate information estimation.

First, we use an exact local expansion of the entropy function to prove almost sure consistency and central limit theorems for several of the most commonly used discretized information estimators. In plain english, this kind of "consistency" theorem says that an estimator "works," in that the estimator is guaranteed to be close to the right value, given enough data; the central limit result sharpens the consistency statement by telling us precisely how much data we need to collect for the estimators to be equal

to the true value, up to a given error tolerance. We also prove tight upper and lower bounds on the bias, or average error, and useful upper bounds on the variance of these estimators.

Second, we prove a converse to the consistency theorems, demonstrating that a misapplication of the most common estimation techniques leads to an arbitrarily poor estimate of the true information, even given unlimited data. This "inconsistency" theorem leads to an analytical approximation of the bias, valid in surprisingly small sample regimes and more accurate than the widely used classical approximation [Miller, 1955, Panzeri et al., 1999] over a large region of parameter space. The two most practical implications of these results are negative: 1) information estimates in a certain data regime are likely contaminated by bias, even if "bias-corrected" estimators were used, and 2) confidence intervals calculated by standard techniques drastically underestimate the error of the most common estimation methods.

Third, we note a very useful connection between the bias of entropy estimators and a certain polynomial approximation problem. By casting bias calculation problems in this approximation theory framework, we obtain the best possible generalization of known asymptotic bias results. We also obtain lower bounds on the convergence rates of any entropy estimator; these bounds give some sense of exactly how difficult this problem is, i.e., how much we can hope to learn about the entropy of any distribution given N samples. More interestingly, this framework leads to an estimator with

some nice properties: the estimator comes equipped with rigorous bounds on the maximum error over all possible underlying probability distributions, and this maximum error turns out to be surprisingly small. We demonstrate the application of this new estimator to both simulated and real data. Matlab code for the computation of this estimator has been placed in the public domain at http://www.cns.nyu.edu/∼liam.

Finally, we collect a number of novel results on, for example: the asymptotic normality of Bayes estimators of entropy (another kind of central limit theorem, but from a completely different point of view) [Wolpert and Wolf, 1995, Nemenman et al., 2002]; the expected error of a certain kind of "approximate sufficiency" analysis, variants of which are currently being developed to search for information-theoretic optimal compressions of the neural code [Victor, 2000b, Gedeon et al., 2003]; permutation symmetry, the existence of symmetric minimax (best in a worst-case sense) estimators of entropy, and the insufficiency of symmetric estimators, all of which provides some justification and background for the main results described above.

This work is to appear in the journal "Neural Computation," and was presented in part at the Society for Neuroscience 2001, Natural Signal Statistics and Neural Coding 2002, and Computational Neuroscience 2002 meetings.

Part 2: Statistical properties of linear-nonlinear cascade models

We analyze the convergence properties of several spike-triggered analysis techniques in the setting of a probabilistic linear-nonlinear (LN) cascade neural encoding model [Brenner et al., 2001, Schwartz et al., 2002, Hunter and Korenberg, 1986]. This model posits that the cell linearly projects the high-dimensional input signal onto some low-dimensional subspace (the "L" step), then fires probabilistically with a rate given by some nonlinear function of this linear projection (the "N" step; see Fig. 0.1). Perhaps the most common example of an LN model is the following caricature of a simple cell in primary visual cortex: the cell (linearly) projects the image onto its receptive field, then spikes with a rate proportional to a (nonlinearly) thresholded value of this inner product operation.

This model is simple and intuitive from a physiological point of view, but presents some interesting statistical challenges. Most of the difficulties can be traced to the fact that the model is "semiparametric" [Begun et al., 1983, van der Vaart, 1998]; one parameter of interest (the linear projection) is finite-dimensional, while the other (the nonlinearity, which is assumed to be completely unknown) is infinite-dimensional. We focus mainly on learning the finite-dimensional parameter, since once the finite-dimensional parameter is known, the infinite-dimensional problem becomes a fairly standard conditional density estimation problem, about which much

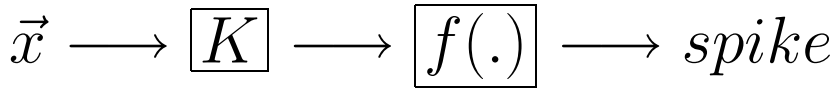

Figure 0.1: Pictorial representation of the LN model. The cell projects the stimulus $\vec{x}$ onto some relatively low-dimensional subspace K, then applies some nonlinearity f to determine the probability of spiking on any given trial. Returning to our V1 simple cell, the stimulus $\vec{x}$ is an image, the linear operator K has rank one and corresponds to an inner product operation between the image $\vec{x}$ and the (roughly Gabor-shaped) receptive field, and f is a thresholding nonlinearity.

is known [Devroye and Lugosi, 2001].

We start by giving exact consistency and central limit (i.e., rate of convergence, as above) results for the common spike-triggered average (STA) technique [Chichilnisky, 2001, Theunissen et al., 2001], which can be viewed as an estimator of this finite-dimensional parameter under certain conditions. Next, we analyze a spike-triggered covariance method, variants of which have been recently exploited successfully by Bialek, Simoncelli, and colleagues [Brenner et al., 2001, Schwartz et al., 2002]; similarly strong consistency and rate-of-convergence results are provided. Unfortunately, the conditions under which these two estimators converge to the correct model parameters in general are quite stringent (for example, these conditions are typically not satisfied by natural signal data or other standard neurophysiological stimulus ensembles).

Therefore, we introduce an estimator for the LN model parameters which

is designed to converge to the true parameter value under general conditions. This estimator is based on the simple idea that not all distinct nonlinearities in the LN model are detectable via their effects on the spike-triggered mean and/or variance; however, there is a class of more general information-theoretic "divergence" functions [Csiszar, 1967] which are able to detect any given nonlinearity, and these divergences can be used to estimate, in turn, the linear projection term. We show that this estimator is consistent in great generality, and derive its rate of convergence. We conclude this mathematical analysis by providing lower bounds on the convergence rate of any possible LN estimator; again, these bounds give some rigorous insight into the difficulty of the LN estimation problem, and also provide an absolute yardstick against which we can measure any candidate estimator [van der Vaart, 1998].

We also develop an efficient, specialized algorithm for the computation of the new estimator; this algorithm makes use of several novel tricks which might be useful more generally for maximizing data-dependent functions on spaces of vector spaces (for example, in "independent components analysis" [Hyvarinen et al., 2001] or "projection pursuit" [Diaconis and Freedman, 1984] applications). We demonstrate the applicability of the algorithm using data from the primary motor cortex of awake behaving monkeys (partially recorded as an undergraduate student in the lab of Dr. John Donoghue, at Brown University, under the guidance of Dr. Nicholas Hatsopoulos and in collaboration with Matthew

Fellows) [Paninski et al., 2003a, Paninski et al., 2002, Shoham et al., 2003]; here, the usual relationship between "stimulus" and "response" is effectively reversed, but the math remains unchanged, and the analysis is therefore applicable without modification. We also give several examples of simulated and real data on which the new estimator outperforms the classical methods (visual neural data from salamander and monkey retina, courtesy of Dr. E.J. Chichilnisky, at the Salk Institute, San Diego) [Schwartz et al., 2002].

These results should prove useful in the study of the neural coding of high-dimensional natural signals, a field which has seen much interest recently [Theunissen et al., 2001, Brenner et al., 2001, Schwartz et al., 2002, Ringach et al., 2002]. For example, we note that [Sharpee et al., 2003] have recently (independently) employed a similar estimator for the estimation of receptive fields from naturalistic data presented to simulated visual and auditory cortical cells. Again, we plan to publicly disseminate Matlab and C (.mex) code for the computation of our estimator at http://www.cns.nyu.edu/∼liam.

Finally, we discuss a few possible extensions of the basic methodology presented here, to neural models which explicitly include the effects of the spiking history of the cell and/or of its neighbors in a multiple-cell network. We consider three such extensions. The first model relaxes the Poisson structure of the spike trains of the basic LN model by allowing a factored (multiplicative) form of refractoriness. We solve a problem posed (implicitly) in a few recent papers [Berry and Meister, 1998,

Aguera y Arcas et al., 2001]: if we know the refractory properties of the cell, how do we recover the linear prefilter? It is not hard to show that spike-triggered averaging and covariance fail in this case, and why; the correct solution turns out to be simple and intuitive.

The second such extension is more specialized but somewhat more grounded in cellular biophysics. We consider an integrate-and-fire cell driven by a linearly-filtered version of the input signal [Reich et al., 1998]; the estimation of the parameters of this model is somewhat more difficult than in the simple LN case. Note, for instance, that while the spike train is a conditional renewal process, given the input signal, the conditional probability of a spike no longer has the factored form present either in the basic LN model or in the LN model with refractoriness introduced in the last paragraph. Thus, nonlinear optimization schemes must be employed to compute the maximum likelihood estimator, which is easily proven to be consistent and statistically efficient for this model. Our main contribution here is that the likelihood surface does not suffer from any of the local minima that often plague nonlinear optimizers, making computation of an efficient estimator tractable [Pillow et al., 2003]. See also [Pillow and Simoncelli, 2003], [Paninski et al., 2003c], and http://www.cns.nyu.edu/∼liam/adapt.html for further analysis of a few of the interesting differences between integrate-and-fire and LN models.

The final extension is perhaps the simplest: we append neural data — either from different cells which might have been recorded simultaneously

on a multielectrode array, or from the same cell at positive leads or lags — to our description of the original signal vector space, and then apply the LN estimation machinery developed here, unchanged, in the new space. We show that, in primate primary motor cortex (MI), observing these side neural effects — the state of the MI network, in some sense — significantly increases the predictability of the firing rate, even given the full kinematic signal. This kind of extension could provide a straightforward but effective way to relax the assumption of conditional independence that plays such a strong role in most applications of Bayesian decoding methods to neural data [Brown et al., 1998, Zhang et al., 1998, Dayan and Abbott, 2001], thus providing a more accurate model of the neural code at a population level.

This work is to appear in the journal "Network: Computation in Neural Systems," for a special issue on the neural coding of natural signals, and was presented in part at the Sloan-Swartz, Computational Neuroscience, Society for Neuroscience, and Neural Information Processing Systems meetings in 2002.

Part 3: Information-theoretic design of experiments

We discuss an idea for collecting data in a relatively efficient manner. Our point of view is Bayesian and information-theoretic: on any given trial, we want to adaptively choose the input in such a way that the mutual informa-

tion between the (unknown) state of the system and the (stochastic) output is maximal, given any prior information (including data collected on any previous trials). We prove a theorem that quantifies the effectiveness of this strategy and give a few illustrative examples comparing the performance of this adaptive technique to the more usual nonadaptive experimental design.

This work is to be presented at the Computational Neuroscience meeting in 2003.

CHAPTER 1

Estimation of entropy and mutual information

1.1 Introduction

The mathematical theory of information transmission represents a pinnacle of statistical research: the ideas are at once beautiful and applicable to a remarkably wide variety of questions. While psychologists and neurophysiologists began to apply these concepts almost immediately after their introduction, the last decade has seen a dramatic increase in the popularity of information-theoretic analysis of neural data. It is unsurprising that these methods have found applications in neuroscience: after all, the theory shows that certain concepts, such as mutual information, are unavoidable when one asks the kind of questions neurophysiologists are interested in. For example, the capacity of an information channel is a fundamental quantity

when one is interested in how much information can be carried by a probabilistic transmission system, such as a synapse. Likewise, we should calculate the mutual information between a spike train and an observable signal in the world when we are interested in asking how much we (or any homunculus) could learn about the signal from the spike train. In this paper, we will be interested not in *why* we should estimate information-theoretic quantities (see, e.g., [Rieke et al., 1997] and [Cover and Thomas, 1991] for extended and eloquent discussions of this question), but rather *how well* we can estimate these quantities at all, given finite i.i.d. data.

One would think this question would be well-understood; after all, applied statisticians have been studying this problem since the first appearance of Shannon's papers, over fifty years ago. Somewhat surprisingly, though, many basic questions have remained unanswered. To understand why, consider the problem of estimating the mutual information, $I(X;Y)$, between two signals X and Y. This estimation problem lies at the heart of the majority of applications of information theory to data analysis; to make the relevance to neuroscience clear, let X be a spike train, or an intracellular voltage trace, and Y some behaviorally relevant, physically observable signal, or the activity of a different neuron. In these examples, and in many other interesting cases, the information estimation problem is effectively infinite-dimensional — by the definition of mutual information, we require knowledge of the joint probability distribution $P(X,Y)$ on the range spaces of X and Y, and these spaces can be quite large — and it would seem to

be very difficult to make progress here in general, given the limited amount of data one can expect to obtain from any physiological preparation.

In this paper, we analyze a discretization procedure for reducing this very hard infinite-dimensional learning problem to a series of more tractable finite-dimensional problems. While we are interested particularly in applications to neuroscience, our results are valid in general, for any information estimation problem. It turns out to be possible to obtain a fairly clear picture of exactly how well the most commonly used discretized information estimators perform, why they fail in certain regimes, and how this performance can be improved. Our main practical conclusions are that the most common estimators can fail badly in common data-analytic situations, and that this failure is more dramatic than has perhaps been appreciated in the literature. The most common procedures for estimating confidence intervals, or "error bars," fail even more dramatically in these "bad" regimes. We suggest a new approach here, and prove some of its advantages.

The paper is organized as follows: in section 1.2, we define the basic regularization procedure (or rather, we formalize an intuitive scheme that has been widely used for decades). We review the known results in section 1.3, then go on in section 1.4 to clarify and improve existing bias and variance results, proving consistency and asymptotic normality results for a few of the most commonly used information estimators. These results serve mainly to show when these common estimators can be expected to be accurate and when they should be expected to break down. The next two sections contain

the central results of this paper: in section 1.5 we show, in a fairly intuitive way, why these common estimators perform so poorly in certain regimes, and exactly how bad this failure is. These results lead us, in section 1.6, to study a polynomial approximation problem associated with the bias of a certain class of entropy estimators; this class includes the most common estimators in the literature, and the solution to this approximation problem provides a new estimator with much better properties. Section 1.7 describes some numerical results that demonstrate the relevance of our analysis for physiological data regimes. We conclude with a brief discussion of three extensions of this work: section 1.8.1 examines a surprising (and possibly useful) degeneracy of a Bayesian estimator, 1.8.2 gives a consistency result for a potentially more powerful regularization method than the one examined in depth here, and 1.8.3 attempts to place our results in the context of estimation of more general functionals of the probability distribution (that is, not just entropy and mutual information). We attach two appendices: in the first, we list a few assorted results which are interesting in their own right but did not fit easily into the flow of the paper; in the second, we give proofs of several of the more difficult results, deferred for clarity's sake from the main body of the text. Throughout, we assume little previous knowledge of information theory beyond an understanding of the definition and basic properties of entropy [Cover and Thomas, 1991]; however, some knowledge of basic statistics is assumed (see, e.g., [Schervish, 1995] for an introduction).

This paper is intended for two audiences: first, applied scientists (especially neurophysiologists) interested in using information-theoretic techniques for data analysis; and second, theorists interested in the more mathematical aspects of the information estimation problem. This split audience could make for a somewhat split presentation: the correct statement of the results requires some mathematical precision, while the demonstration of their utility requires some more verbose explanation. Nevertheless, we feel that the intersection between the applied and theoretical communities is large enough to justify a unified presentation of our results and our motivations; we hope the reader will agree, and forgive the length of the resulting manuscript.

1.2 The setup: Grenander's method of sieves

As discussed above, much of the inherent difficulty of our estimation problem stems from the fact that the mutual information,

$$I(X,Y) \equiv \int_{\mathcal{X}\times\mathcal{Y}} dP(x,y) \log \frac{dP(x,y)}{d(P(x)\times P(y))}$$

is a nonlinear functional of an unknown joint probability measure, $P(X,Y)$, on two arbitrary measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. In many interesting cases, the "parameter space" — the space of probability measures under consideration — can be very large, even infinite-dimensional. For example, in the neuroscientific data analysis applications that inspired this work [Strong et al., 1998], $\mathcal{X}$ could be a space of time-varying visual stimuli, and

$\mathcal{Y}$ the space of spike trains that might be evoked by a given stimulus; this $\mathcal{Y}$ could be taken to be a (quite large) space of discrete (counting) measures on the line, while $\mathcal{X}$ could be modeled as the (even larger) space of generalized functions on $\Re^3$. Given N i.i.d. samples from $P(X,Y)$, $\{x_i, y_i\}_{1 \leq i \leq N}$, ("stimulus" together with the evoked "response"), how well can we estimate the information this cell provides the brain about the visual scene? Clearly, it is difficult to answer this question as posed; the relationship between stimulus and response could be too complex to be revealed by the available data, even if N is large by neurophysiological standards. In fact, there are general theorems to this effect (section 1.3). Therefore, some kind of regularization is needed.

The most successful approach taken to date in our field to circumvent these problems was introduced by Bialek and colleagues [Bialek et al., 1991, Strong et al., 1998]. The idea is to admit to the difficulty of the problem, and instead estimate a system of lower bounds on the mutual information via the data processing inequality [Cover and Thomas, 1991], which states that

$$I(X;Y) \geq I(S(X);T(Y)),$$

for any random variables X and Y and any functions S and T on the range of X and Y, respectively. The generality of the data processing inequality implies that we are completely unconstrained in our choice of S and T. So the strategy, roughly, is to choose a sequence of functions S_N and T_N

which preserve as much information as possible given that $I(S_N; T_N)$ can be estimated with some fixed accuracy from N data samples. (Note that S_N and T_N are chosen independently of the data.) As the size of the available data set increases, our lower bound grows monotonically towards the true information. In slightly different language, S_N and T_N could be viewed as models, or parametrizations of the allowed underlying measures $P(X, Y)$; we are simply allowing our model to become richer (higher-dimensional) as more data becomes available for fitting. Clearly, then, we are not introducing anything particularly novel, but merely formalizing what statisticians have been doing naturally since well before Shannon had written his papers.

This strategy bears a striking resemblance to regularization methods employed in abstract statistical inference [Grenander, 1981], generally known as the "method of sieves." Here one replaces the parameter space of interest with a closely related space which simplifies the analysis, or provides estimators with more attractive statistical properties. The following example is canonical and helps to clarify exactly why regularization is necessary. Say one is sampling from some unknown, smooth probability density function, and one is interested in estimating the underlying density. It is clear that there exists no maximum likelihood estimator of the density in the space of smooth functions (the object which formally maximizes the likelihood, a sum of Dirac point masses, does not lie in the allowed smoothness class). The situation is pathological, then: as the sample size increases to infinity, our estimate does not converge to the true density in the sense of any

smooth topology. To avoid this pathology, we regularize our estimator by requiring that it take its values in a smooth function space. In effect, we restrict our attention to a subset, a "sieve," of the possible parameter space. As the available data increase, we gradually relax our constraints on the smoothness of the estimator (decrease the "mesh size" of our sieve), until in the limit our estimate of the underlying density is almost surely arbitrarily close to the true density. We will borrow this "mesh" and "sieve" terminology for the remainder of the paper.

Here, we have to estimate a joint probability measure, $P(X, Y)$, on a large product space, $\mathcal{X} \times \mathcal{Y}$, in order to compute $I(X;Y)$. This is very difficult; therefore, we regularize our problem by instead trying to estimate $P(S, T)$ (where $P(S, T)$ is induced by the maps S and T in the natural way, i.e., $P(S = i, T = j) = P((x, y) : S(x) = i, T(y) = j)$). Thus our "mesh size" is determined by the degree of compression inherent in going from (x, y) to $(S(x), T(y))$. Two variants of this strategy have appeared in the neuroscientific literature. The first, the so-called "reconstruction" technique [Bialek et al., 1991], makes use of some extremal property of the prior signal distribution to facilitate the reliable estimation of a lower bound on the true information. T_N here is a series of convolution operators, mapping spike trains (elements of $\mathcal{Y}$) back into the signal space $\mathcal{X}$. The lower bound on the information $I(X, T_N(Y))$ is estimated by spectral techniques: the prior distribution of X, $P(X)$, is chosen to be Gaussian, and the well-known maximum-entropy property and spectral information formula for Gaussian

distributions provide the desired bound. The lower bounds obtained by this reconstruction approach have proven quite useful [Rieke et al., 1997]; however, the available convergence results (of $I(X, T_N(Y))$ to $I(X, Y)$ as $N \to \infty$) rely on strong assumptions on $P(X, Y)$, and we will not discuss this technique in depth. (One final note: the reader familiar with the reconstruction technique will realize that this example does not quite fit into our general framework, as the convolution operators T_N, which are chosen via regression techniques, are in fact dependent on the data. These dependencies complicate the analysis significantly, and we will say very little on this topic beyond a brief note in section 1.8.2.)

The second method, the so-called "direct method," [Strong et al., 1998, Buracas et al., 1998] is at first sight less dependent on assumptions on the prior distribtion on $\mathcal{X}$. Here one discretizes the space of all spike trains on some interval $[0, T]$ into some finite number, m, of words w, and makes use of the information formula for discrete distributions,

$$I(X; W) = H(W) - H(W|X),$$

to obtain a lower bound on the mutual information between the spike train and the signal of interest. $H(.)$ above denotes the entropy functional,

$$H(W) \equiv -\sum_i P(W_i) \log P(W_i),$$

and $H(.|.)$ denotes conditional entropy; X is say, a visual signal on which we are conditioning.[1] In our previous notation, $W(y) = T(y)$. The generality of

[1]We should note that, to keep data requirements manageable, $H(W|X)$ —

the data processing inequality, again, means that the discretization can take arbitrary form; letting T depend on the data size N, T_N could, for example, encode the total number of spikes emitted by the neuron for small N, then the occurrence of more detailed patterns of firing [Strong et al., 1998] for larger N, until, in the limit, all of the information in the spike train is retained.

Thus, in this "direct" approach, S_N and T_N are as simple as possible: these maps discretize $\mathcal{X}$ and $\mathcal{Y}$ into a finite number of points, $m_{S,N}$ and $m_{T,N}$, where $m_{S,N}$ and $m_{T,N}$ grow with N. For each value of N, our problem reduces to estimating $I(S_N, T_N)$, where the joint distribution of the random variables S_N and T_N is discrete on $m_{S,N}m_{T,N}$ points, and our parameter space, far from being infinite-dimensional, is the tractable $m_{S,N}m_{T,N}$-simplex, the set of convex combinations of $m_{S,N}m_{T,N}$ disjoint point masses. We emphasize again that neither S, T, nor m are allowed to depend on the data; in effect, we pretend that the discretizing maps and their ranges are chosen in advance, before we see a single sample.

While this discrete "binning" approach appears quite crude, it will allow us to state completely general strong convergence theorems for the

the expected conditional entropy of W given x, averaged over $P(X)$ — is often replaced with $H(W|x)$, the conditional entropy given only a single x. The fact that any rigorous justification of this substitution requires a strong assumption (namely, that $H(W|x)$ is effectively independent of x with high $P(x)$-probability) has perhaps been overly glossed over in the literature.

information estimation problem, without any assumptions on, say, the existence or smoothness of a density for $P(X, Y)$. To our knowledge, results of this generality are unavailable outside the discrete context (but see [Beirlant et al., 1997] for a good review of differential entropy estimation techniques, which provide a powerful alternative approach when the underlying probability measures are known *a priori* to possess a given degree of smoothness [Victor, 2002]). In addition, of course, data which naturally take only a finite number of values are not uncommon. Therefore, we will analyze this discrete approach exclusively for the remainder of this paper.

1.3 Previous work

Most of the following results are stated in terms of the entropy $H(X)$; corresponding results for $I(X, Y)$ follow by Shannon's formula for discrete information:

$$I(X, Y) = H(X) + H(Y) - H(X, Y).$$

All of the estimators we will consider are functionals of the "empirical measures"

$$p_{N,i} \equiv \frac{1}{N} \sum_{j=1}^{N} \delta_i(T_N(y_j))$$

(where δ_i denotes the probability measure concentrated at i). The three most popular estimators for entropy seem to be:

1. The maximum likelihood (ML) estimator given p_N (also called the

"plug-in" [Antos and Kontoyiannis, 2001] or "naive" [Strong et al., 1998] estimator),

$$\hat{H}_{MLE}(p_N) \equiv -\sum_{i=1}^{m} p_{N,i} \log p_{N,i}$$

(all logs are natural unless stated otherwise);

2. The MLE with the so-called Miller-Madow bias correction [Miller, 1955],

$$\hat{H}_{MM}(p_N) \equiv \hat{H}_{MLE}(p_N) + \frac{\hat{m}-1}{2N},$$

where $\hat{m}$ is some estimate of the number of bins with non-zero P-probability (here we take $\hat{m}$ to be the number of bins with nonzero p_N-probability; see [Panzeri and Treves, 1996] for some other examples);

3. The jackknifed [Efron and Stein, 1981] version of the MLE,

$$\hat{H}_{JK} \equiv N\hat{H}_{MLE} - \frac{N-1}{N}\sum_{j=1}^{N} \hat{H}_{MLE-j},$$

where H_{MLE-j} is the MLE based on all but the j-th sample (unpublished notes of J. Victor; see also, e.g., [Strong et al., 1998], in which a very similar estimator is used).

1.3.1 CLT, asymptotic bias and variance

The majority of known results are stated in the following context: fix some discrete measure p on m bins and let N tend to infinity. In this case,

the multinomial central limit theorem implies that the empirical measures p_N are asymptotically normal, concentrated on an ellipse of size $\sim N^{-1/2}$ around the true discrete measure p; since $\hat{H}_{MLE}$ is a smooth function of p on the interior of the m-simplex, $\hat{H}_{MLE}$ is asymptotically normal (or chi-squared or degenerate, according to the usual conditions [Schervish, 1995]) as well. It follows that both the bias and variance of $\hat{H}_{MLE}$ decrease approximately as $\frac{1}{N}$ [Basharin, 1959] at all but a finite number of points on the m-simplex. We will discuss this bias and variance rate explicitly for the above estimators in section 1.4; here it is sufficient to note that the asymptotic variance rate varies smoothly across the space of underlying probability measures $p(x, y)$, while the bias rate depends only on the number of nonzero elements of p (and is therefore constant on the interior of the m-simplex and discontinuous on the boundary). The asymptotic behavior of this estimation problem (again, when m is fixed and $N \to \infty$) is thus easily handled by classical techniques. While it does not seem to have been noted previously, it follows from the above that $\hat{H}_{MLE}$ is asymptotically minimax for fixed m as $N \to \infty$ (by "minimax" we mean best in a worst-case sense; we will discuss this concept in more detail below); see, e.g., [Prakasa Rao, 2001] for the standard technique, a clever "local Bayesian" application of the Cramer-Rao inequality.

There have also been several papers [Miller, 1955, Carlton, 1969, Treves and Panzeri, 1995, Victor, 2000a] providing a series expansion for the bias, in the hope of estimating and subtracting out the bias directly.

While these authors have all arrived at basically the same answer, they have done so with varying degrees of rigor: for example, [Miller, 1955] use an expansion of the logarithm which is not everywhere convergent (we ouline this approach below and show how to avoid these convergence problems). [Carlton, 1969] rearranged the terms of a convergent expansion of the logarithm term in H; unfortunately, this expansion is not absolutely convergent, and therefore this rearrangement is not necessarily justified. [Treves and Panzeri, 1995] and [Victor, 2000a] both admit that their methods (a divergent expansion of the logarithm in each case) are not rigorous. Therefore, it would appear that none of the available results are strong enough to use in the context of this paper, where m and p can depend arbitrarily strongly on N. We will remedy this situation below.

1.3.2 Results of Antos and Kontoyiannis, '01

[Antos and Kontoyiannis, 2001] recently contributed two relevant results. The first is somewhat negative:

Theorem ([Antos and Kontoyiannis, 2001]). *For any sequence $\{\hat{H}_N\}$ of entropy estimators, and for any sequence $\{a_N\}$, $a_N \searrow 0$, there is a distribution P on the integers $\mathcal{Z}$ with $H \equiv H(P) < \infty$ and*

$$\limsup_{n\to\infty} \frac{E(|\hat{H}_N - H|)}{a_N} = \infty.$$

In other words, there is no universal rate at which the error goes to zero, no matter what estimator we pick, even when our sample space is discrete

(albeit infinite); given any such putative rate a_N, we can always find some distribution P for which the true rate of convergence is infinitely slower than a_N. [Antos and Kontoyiannis, 2001] prove identical theorems for the mutual information, as well as a few other functionals of P.

The second result is an easy consequence of a more general fact about functions of multiple random variables; since we will use this general theorem repeatedly below, we reproduce the statement here. See, e.g., [McDiarmid, 1989, Devroye et al., 1996] for a proof and extended discussions; the result basically says that if f is a function of N independent random variables, such that f depends only weakly on the value of any single variable, then f is tightly concentrated about its mean (i.e., $Var(f)$ is small).

Theorem ("McDiarmid's inequality"; Chernoff, Azuma et al.). *If $\{x_j\}_{j:1,\ldots,N}$ are independent random variables taking values in some arbitrary measurable space A, and $f : A^N \mapsto \Re$ is some function satisfying the coordinatewise boundedness condition*

$$\sup_{\{x_1,\ldots,x_N\},x'_j} |f(x_1,\ldots,x_N) - f(x_1,\ldots,x_{j-1},x'_j,x_{j+1},\ldots,x_N)| < c_j,\ 1 \leq j \leq N, \tag{1.1}$$

then, for any $\epsilon > 0$,

$$P(|f(x_1,\ldots,x_N) - E(f(x_1,\ldots,x_N))| > \epsilon) \leq 2e^{-2\epsilon^2/\sum_{j=1}^N c_j^2}. \tag{1.2}$$

The condition says that, by changing the value of the coordinate x_j, we can not change the value of the function f by more than some constant c_j.

The usefulness of the theorem is a result both of the ubiquity of functions f satifying condition (1.1) (and the ease with which we can usually check the condition), and of the exponential nature of the inequality, which can be quite powerful if $\sum_{j=1}^{N} c_j^2$ satisfies reasonable growth conditions.

Antos and Kontoyiannis [Antos and Kontoyiannis, 2001] pointed out that this leads easily to a useful bound on the variance of the MLE for entropy:

Theorem (Antos and Kontoyiannis, '01). *a. For all N, the variance of the MLE for entropy is bounded above:*

$$Var(\hat{H}_{MLE}) \leq (\frac{(\log N)^2}{N}). \tag{1.3}$$

b. Moreover, by McDiarmid's inequality (1.2),

$$P(|\hat{H}_{MLE} - E(\hat{H}_{MLE})| > \epsilon) \leq 2e^{\frac{-N}{2}\epsilon^2(\log N)^{-2}}. \tag{1.4}$$

Note that, although this inequality is not particularly tight — while it says that the variance of $\hat{H}_{MLE}$ necessarily dives to zero with increasing N, the true variance turns out to be even smaller than the bound indicates — the inequality is completely universal, i.e., independent of m or P. For example, [Antos and Kontoyiannis, 2001] use it in the context of m (countably) infinite. In addition, it is easy to apply this result to other functionals of p_N; see section 1.6 for one such important generalization.

1.3.3 $\hat{H}_{MLE}$ is negatively biased everywhere

Finally, for completeness, we mention the following well-known fact:

$$E_p(\hat{H}_{MLE}) \leq H(p), \tag{1.5}$$

where $E_p(.)$ denotes the conditional expectation given p. We have equality in the above expression only when $H(p) = 0$; in words, the bias of the MLE for entropy is negative everywhere unless the underlying distribution p is supported on a single point. This is all a simple consequence of Jensen's inequality; a proof was recently given in [Antos and Kontoyiannis, 2001], and we will supply another easy proof below. Note that (1.5) does *not* imply that the MLE for mutual information is biased upwards everywhere, as has been claimed elsewhere; it is easy to find distributions p such that $E_p(\hat{I}_{MLE}) < I(p)$. We will discuss the reason for this misunderstanding below.

It will help to keep the following Figure 1.1 in mind. This figure gives a compelling illustration of perhaps the most basic fact about $\hat{H}_{MLE}$: the variance is small and the bias is large until $N >> m$. This qualitative statement is not new; however, the corresponding quantitative statement — especially the fact that $\hat{H}_{MLE}$ in the statement can be replaced with any of the three most commonly used estimators — appears to be novel. We will develop this argument over the next four sections, and will postpone our discussion of the implications for data analysis until the conclusion.

1.4 The $N >> m$ range: the local expansion

The unifying theme of this section is a simple local expansion of the entropy functional around the true value of the discrete measure p, a variant of what is termed the "delta method" in the statistics literature. This expansion is similar to one used by previous authors; we will be careful to note the extensions provided by the current work.

The main idea, outlined, e.g., in [Serfling, 1980], is that any smooth function of the empirical measures p_N (e.g., any of the three estimators for entropy introduced above) will behave like an affine function with probability approaching one as N goes to infinity. To be more precise, given some functional f of the empirical measures, we can expand f around the underlying distribution p as follows:

$$f(p_N) = f(p) + df(p; p_N - p) + r_N(f, p, p_N),$$

where $df(p; p_N - p)$ denotes the "functional derivative" (Frechet derivative) of f with respect to p in the direction $p_N - p$, and $r_N(f, p, p_N)$ the remainder. If f is sufficiently smooth (in a suitable sense), the differential $df(p; p_N - p)$ will be a linear functional of $p_N - p$ for all p, implying

$$df(p; p_N - p) \equiv df(p; \frac{1}{N}\sum_{j=1}^{N} \delta_j - p) = \frac{1}{N}\sum_j df(p; \delta_j - p),$$

i.e., $df(p; p_N - p)$ is the average of N i.i.d variables, which implies, under classical conditions on the tail of the distribution of $df(p; \delta_j - p)$, that $N^{1/2} df(p; p_N - p)$ is asymptotically normal. If we can prove that

$N^{1/2} r_N(f, p, p_N)$ goes to zero in probability (that is, the behavior of f is asymptotically the same as the behavior of a linear expansion of f about p), then a central limit theorem for f follows. This provides us with a more flexible approach than the method outlined in section 1.3.1 (recall that that method relied on a CLT for the underlying empirical measures p_N, and such a CLT does not necessarily hold if m and p are not fixed).

Let us apply all this to H:

$$\begin{aligned} \hat{H}_{MLE}(p_N) &= H(p_N) \\ &= H(p) + dH(p; p_N - p) + r_N(H, p, p_N) \\ &= H(p) + \sum_{i=1}^{m} (p_i - p_{N,i}) \log p_i + r_N(H, p, p_N). \end{aligned} \tag{1.6}$$

A little algebra shows that

$$r_N(H, p, p_N) = -D_{KL}(p_N; p),$$

where $D_{KL}(p_N; p)$ denotes the Kullback-Leibler divergence between p_N, the empirical measure, and p, the true distribution. The sum in (1.6) has mean 0; by linearity of expectation, then,

$$E_p(\hat{H}_{MLE}) - H = -E_p(D_{KL}(p_N; p)), \tag{1.7}$$

and since $D_{KL}(p_N; p) \geq 0$, where the inequality is strict with positive probability whenever p is nondegenerate, we have a simple proof of the nonpositive bias of the MLE. Another (slightly more informative) approach will be given in section 1.6.

The second useful consequence of the local expansion follows by the next two well-known results [Gibbs and Su, 2002]:

$$0 \leq D_{KL}(p_N; p) \leq \log(1 + \chi^2(p_N; p)), \tag{1.8}$$

where

$$\chi^2 \equiv \sum_{i=1}^{m} \frac{(p_{N,i} - p_i)^2}{p_i^2}$$

denotes Pearson's chi-square functional, and

$$E_p(\chi^2(p_N; p)) = \frac{|\text{supp}(p)| - 1}{N} \ \forall p, \tag{1.9}$$

where $|\text{supp}(p)|$ denotes the size of the support of p, the number of points with nonzero p-probability. Expressions 1.7, 1.8, and 1.9, with Jensen's inequality, give us rigorous upper and lower bounds on $B(\hat{H}_{MLE})$, the bias of the MLE:

Proposition 1.

$$-\log(1 + \frac{m-1}{N}) \leq B(\hat{H}_{MLE}) \leq 0,$$

with equality iff p is degenerate. The lower bound is tight as $N/m \to 0$, and the upper bound is tight as $N/m \to \infty$.

Here we note that [Miller, 1955] used a similar expansion to obtain the $\frac{1}{N}$ bias rate for m fixed, $N \to \infty$. The remaining step is to expand $D_{KL}(p_N; p)$:

$$D_{KL}(p_N; p) = \frac{1}{2}(\chi^2(p_N; p)) + O(N^{-2}), \tag{1.10}$$

if p is fixed. As noted in section 1.3.1, this expansion of D_{KL} does not converge for all possible values of p_N; however, when m and p are fixed, it is easy to show, using a simple cutoff argument, that this "bad" set of p_N has an asymptotically negligible effect on $E_p(D_{KL})$. The formula for the mean of the chi-square statistic (equation (1.9) above) completes Miller's and Madow's original proof; we have [Miller, 1955]

$$B(\hat{H}_{MLE}) = -\frac{m-1}{2N} + o(N^{-1}), \tag{1.11}$$

if m is fixed and $N \to \infty$. From here it easily follows that $\hat{H}_{MM}$ and $\hat{H}_{JK}$ both have $o(N^{-1})$ bias under these conditions (for $\hat{H}_{MM}$, we need only show that $\hat{m} \to m$ sufficiently rapidly, and this follows by any of a number of exponential inequalities [Dembo and Zeitouni, 1993, Devroye et al., 1996]; the statement for $\hat{H}_{JK}$ can be proven by direct computation). To extend these kinds of results to the case when m and p are not fixed, we have to generalize (1.11); this desired generalization of Miller's result does turn out to be true, as we prove (using a completely different technique) in section 1.6.

It is worth emphasizing that $E_p(\chi^2(p_N; p))$ is *not* constant in p; it is constant on the interior of the m-simplex, but varies discontinuously on the boundary. This was the source of the confusion about the bias of the MLE for information,

$$\hat{I}_{MLE}(x, y) \equiv \hat{H}_{MLE}(x) + \hat{H}_{MLE}(y) - \hat{H}_{MLE}(x, y);$$

when $p(x, y)$ has support on the full $m_x m_y$ points, the $\frac{1}{N}$ bias rate is indeed

given by $m_x m_y - m_x - m_y - 1$, which is positive for m_x, m_y large enough. However, $p(x, y)$ can be supported on as few as $\max(m_x, m_y)$ points, which means that the $\frac{1}{N}$ bias rate of $\hat{I}_{MLE}$ can be negative. It could be argued that this reduced-support case is non-physiological; however, a simple continuity argument shows that even when $p(x, y)$ has full support but places most of its mass on a subset of its support, the bias can be negative even for large N, even though the asymptotic bias rate in this case is positive.

The simple bounds of Proposition 1 form about half of the proof of the following two theorems, the main results of this section: they say that if $m_{S,N}$ and $m_{T,N}$ grow with N, but not too quickly, the "sieve" regularization works, in the sense that the sieve estimator is almost surely consistent and asymptotically normal and efficient on a $\sqrt{N}$ scale. The power of these results lie in their complete generality: we place no constraints whatsoever on either the underlying probability measure, $p(x, y)$, or the sample spaces $\mathcal{X}$ and $\mathcal{Y}$. Note that the theorems are true for all three of the estimators defined above (i.e., $\hat{H}$ above — and in the rest of the paper, unless otherwise noted — can be replaced by $\hat{H}_{MLE}$, $\hat{H}_{JK}$, or $\hat{H}_{MM}$); thus, all three common estimators have the same $\frac{1}{N}$ variance rate: σ^2, as defined below. In the following, $\sigma_{X,Y}$ is the joint σ-algebra of $X \times Y$ on which the underlying probability distribution $p(X, Y)$ is defined, σ_{S_N,T_N} is the (finite) σ-algebra generated by S_N and T_N, and H_N denotes the N-discretized entropy, $H(S_N(X))$. The σ-algebra condition in Theorem 2 is merely a technical way of saying that S_N and T_N asymptotically retain all of the

data in the sample (x, y) in the appropriate measure-theoretic sense; see the appendix for details.

Theorem 2 (Consistency). *If* $m_{S,N} m_{T,N} = o(N)$ *and* σ_{S_N,T_N} *generates* $\sigma_{X,Y}$, *then* $\hat{I} \to I$ *a.s. as* $N \to \infty$.

Theorem 3 (Central limit). *Let*

$$\sigma_N^2 \equiv Var(-\log p_{T_N}) \equiv \sum_{i=1}^{m} p_{T_N,i}(-\log p_{T_N,i} - H_N)^2.$$

If $m_N \equiv m = o(N^{1/2})$, *and*

$$\liminf_{N\to\infty} N^{1-\alpha}\sigma_N^2 > 0$$

for some $\alpha > 0$, *then* $(\frac{N}{\sigma_N^2})^{1/2}(\hat{H} - H_N)$ *is asymptotically standard normal.*

The following lemma is the key to the proof of Theorem 2, and is interesting in its own right:

Lemma 4. *If* $m = o(N)$, *then* $\hat{H} \to H_N$ *a.s.*

Note that σ_N^2 in the statement of the CLT (Theorem 3) is exactly the variance of the sum in expression (1.6), and corresponds to the asymptotic variance derived originally in [Basharin, 1959] (by a similar local expansion). We also point out that σ_N^2 has a specific meaning in the theory of data compression (where σ_N^2 goes by the name of "minimal coding variance"); see [Kontoyiannis, 1997] for more details.

We close this section with some useful results on the variance of $\hat{H}$. We have, under the stated conditions, that the variance of $\hat{H}$ is of order $\frac{\sigma^2}{N}$

asymptotically (by the CLT), and strictly less than $\frac{Clog(N)^2}{N}$ for all N, for some fixed C (by the result of [Antos and Kontoyiannis, 2001]). It turns out that we can "interpolate," in a sense, between the (asymptotically loose but good for all N) p-independent bound and the (asymptotically exact but bad for small N) p-dependent Gaussian approximation. The trick is to bound the average fluctuations in $\hat{H}$ when randomly replacing one sample, instead of the worst-case fluctuations, as in McDiarmid's bound. The key inequality is due to Steele [Steele, 1986]:

Theorem (Steele's inequality). *If $S(x_1, x_2, \ldots, x_N)$ is any function of N i.i.d. random variables then*

$$var(S) \leq \frac{1}{2} E \sum_{j=1}^{N} (S - S_i)^2,$$

where $S_i = S(x_1, x_2, \ldots, x_i', \ldots, x_N)$ is given by replacing the x_i with an i.i.d. copy.

For $S = \hat{H}$, it turns out to be possible to compute the right-hand side explicitly; the details are given in the appendix. It should be clear even without any computation that the bound so obtained is at least as good as the $\frac{Clog(N)^2}{N}$ guaranteed by McDiarmid; it is also easy to show, by the linear expansion technique employed above, that the bound is asymptotically tight under conditions similar to those of Theorem 3.

Thus $\sigma(p)^2$ plays the key role in determining the variance of $\hat{H}$. We know σ^2 can be zero for some p, since $Var(-\log p_i)$ is zero for any p uniform on

any k points, $k \leq m$. On the other hand, how large can σ^2 be? The following proposition provides the answer; the proof is in the appendix.

Proposition 5.

$$\max_p \sigma^2 \sim (\log m)^2.$$

This leads us to define the following bias-variance balance function, valid in the $N >> m$ range:

$$V/B^2 \approx \frac{N(\log m)^2}{m^2};$$

if V/B^2 is large, variance dominates the mean-square error (in the "worst-case" sense), and bias dominates if V/B^2 is small. It is not hard to see that if m is at all large, bias dominates until N is relatively huge (recall Figure 1.1). (This is just a rule of thumb, of course, not least because the level of accuracy desired, and the relative importance of bias and variance, depend on the application. We give more precise — in particular, valid for all values of N and m — formulae for the bias and variance in the following.)

To summarize, the sieve method is effective and the asymptotic behavior of $\hat{H}$ is well understood for $N >> m$. In this regime, if $V/B^2 > 1$, classical (Cramer-Rao) effects dominate, and the three most common estimators ($\hat{H}_{MLE}$, $\hat{H}_{MM}$, and $\hat{H}_{JK}$) are approximately equivalent, since they share the same asymptotic variance rate. On the other hand, if $V/B^2 < 1$, bias plays a more important role and estimators which are specifically designed to reduce the bias become competitive; previous work has demonstrated that $\hat{H}_{MM}$ and $\hat{H}_{JK}$ are effective in this regime [Panzeri and Treves, 1996,

Strong et al., 1998]. We turn in the next section to a regime which is much more poorly understood, the (not uncommon) case when $N \sim m$. We will see that the local expansion becomes much less useful in this regime, and a different kind of analysis is required.

1.5 The $N \sim m$ range: consequences of symmetry

The main result of this section is as follows: if N/m is bounded, the bias of $\hat{H}$ remains large while the variance is always small, even if $N \to \infty$. The basic idea is that entropy is a symmetric function of $p_i, 1 \leq i \leq m$, in that H is invariant under permutations of the points $\{1, \ldots, m\}$. Most common estimators of H, including $\hat{H}_{MLE}$, $\hat{H}_{MM}$, and $\hat{H}_{JK}$, share this permutation symmetry (in fact, one can show that there is some statistical justification for restricting our attention to this class of symmetric estimators; see the appendix). Thus, the distribution of $\hat{H}_{MLE}(p_N)$, say, is the same as that of $\hat{H}_{MLE}(p'_N)$, where p'_N is the rank-sorted empirical measure (for concreteness, define "rank-sorted" as "rank-sorted in decreasing order"). This leads us to study the limiting distribution of these sorted empirical measures (Fig. 1.2). It turns out that these sorted histograms converge to the "wrong" distribution under certain circumstances. We have the following result:

Theorem 6 (Convergence of sorted empirical measures; inconsistency). *Let P be absolutely continuous with respect to Lebesgue measure*

on the interval $[0,1]$*, and let* $p = dP/dm$ *be the corresponding density. Let* S_N *be the* m*-equipartition of* $[0,1]$*,* p' *denote the sorted empirical measure, and* $N/m \to c,\ 0 < c < \infty$*. Then:*

a) $p' \overset{L_1,a.s.}{\to} p'_{c,\infty}$*, with* $||p'_{c,\infty} - p||_1 > 0$*. Here* $p'_{c,\infty}$ *is the monotonically decreasing step density with gaps between steps* j *and* $j+1$ *given by*

$$\int_0^1 dt e^{-cp(t)} \frac{(cp(t))^j}{j!}.$$

b) Assume p *is bounded. Then* $\hat{H} - H_N \to B_{c,\hat{H}}(p)$ *a.s., where* $B_{c,\hat{H}}(p)$ *is a deterministic function, nonconstant in* p*. For* $\hat{H} = \hat{H}_{MLE}$*,*

$$B_{c,\hat{H}}(p) = h(p') - h(p) < 0,$$

where $h(.)$ *denotes differential entropy.*

In other words, when the sieve is too fine ($N \sim m$), the limit sorted empirical histogram exists (and is surprisingly easy to compute) but is not equal to the true density, even when the original density is monotonically decreasing and of step form. As a consequence, $\hat{H}$ remains biased even as $N \to \infty$. This in turn leads to a strictly positive lower bound on the asymptotic error of $\hat{H}$ over a large portion of the parameter space. The basic phenomenon is illustrated in Figure 1.2.

We can apply this theorem to obtain simple formulae for the asymptotic bias $B(p,c)$ for special cases of p: for example, for the uniform distribution $U \equiv U([0,1])$,

$$B_{c,\hat{H}_{MLE}}(U) = \log(c) - e^{-c} \sum_{j=1}^{\infty} \frac{c^{j-1}}{(j-1)!} \log(j);$$

$$B_{c,\hat{H}_{MM}}(U) = B_{c,\hat{H}_{MLE}}(U) + \frac{1-e^{-c}}{2c};$$

$$B_{c,\hat{H}_{JK}}(U) = 1 + \log(c) - e^{-c}\sum_{j=1}^{\infty}\frac{c^{j-1}}{(j-1)!}(j-c)\log(j).$$

To give some intuition on these formulae, note that $B_{c,\hat{H}_{MLE}}(U)$ behaves like $\log(N) - \log(m)$ as $c \to 0$, as expected given that $\hat{H}_{MLE}$ is supported on $[0, \log(N)]$ (recall the lower bound of Proposition 1); meanwhile,

$$B_{c,\hat{H}_{MM}}(U) \sim B_{c,\hat{H}_{MLE}}(U) + \frac{1}{2}$$

and

$$B_{c,\hat{H}_{JK}}(U) \sim B_{c,\hat{H}_{MLE}}(U) + 1$$

in this $c \to 0$ limit. In other words, in the extremely undersampled limit, the Miller correction reduces the bias by only half a nat, while the jackknife only gives us twice that. It turns out that the proof of the theorem leads to good upper bounds on the approximation error of these formulae, indicating that these asymptotic results will be useful even for small N. We examine the quality of these approximations for finite N and m in section 1.7.

This asymptotically deterministic behavior of the sorted histograms is perhaps surprising, given that there is no such corresponding deterministic behavior for the unsorted histograms (although, by the Glivenko-Cantelli theorem [van der Vaart and Wellner, 1996], there is well-known deterministic behavior for the integrals of the histograms). What is going on here? In crude terms, the sorting procedure "averages over" the variability in the unsorted histograms. In the case of the theorem, the "variability" at each

bin turns out to be of a Poisson nature, in the limit as $m, N \to \infty$, and this leads to a well-defined and easy-to-compute limit for the sorted histograms.

To be more precise, note that the value of the sorted histogram at bin k is greater than t if and only if the number of (unsorted) $p_{N,i}$ with $p_{N,i} > t$ is at least k (remember that we are sorting in decreasing order). In other words,

$$p'_N = F_N^{-1},$$

where F_N is the empirical "histogram distribution function,"

$$F_N(t) \equiv \frac{1}{m} \sum_{i=1}^{m} 1(p_{N,i} < t),$$

and its inverse is defined in the usual way. We can expect these sums of indicators to converge to the sums of their expectations, which in this case are given by

$$E(F_N(t)) = \frac{1}{m} \sum_{i} P(p_{N,i} < t);$$

finally, it is not hard to show that this last sum can be approximated by an integral of Poisson probabilites (see appendix for details). Something similar happens even if $m = o(N)$; in this case, under similar conditions on p, we would expect each $p_{N,i}$ to be approximately Gaussian, instead of Poisson.

To compute $E(\hat{H})$ now, we only need note the following important fact: each $\hat{H}$ is a linear functional of the "histogram order statistics"

$$h_j \equiv \sum_{i=1}^{m} 1(n_i = j),$$

where

$$n_i \equiv N p_{N,i}$$

is the unnormalized empirical measure. For example,

$$\hat{H}_{MLE} = \sum_{j=0}^{N} a_{\hat{H}_{MLE},j,N} h_j,$$

where

$$a_{\hat{H}_{MLE},j,N} = -\frac{j}{N} \log \frac{j}{N},$$

while

$$a_{\hat{H}_{JK},j,N} = N a_{\hat{H}_{MLE},j,N} - \frac{N-1}{N}\left((N-j) a_{\hat{H}_{MLE},j,N-1} + j a_{\hat{H}_{MLE},j-1,N-1}\right).$$

Linearity of expectation now makes things very easy for us:

$$\begin{aligned} E(\hat{H}) &= \sum_{j=0}^{N} a_{\hat{H},j,N} E(h_j) \\ &= \sum_{j} \sum_{i=1}^{m} a_{j,N} P(n_i = j) \\ &= \sum_{j} a_{j,N} \sum_{i} \binom{N}{j} p_i^j (1-p_i)^{N-j}. \end{aligned} \tag{1.12}$$

We emphasize that the above formula is exact, for all N, m, and p; again, the usual Poisson or Gaussian approximations to the last sum lead to useful asymptotic bias formulae. See the appendix for the rigorous computations.

For our final result of this section, let p and S_N be as in the statement of Theorem 6, with p bounded, and $N = O(m^{1-\alpha})$, $\alpha > 0$. Then some easy computations show that $P(\exists i : n_i > j) \to 0$ for all $j > \alpha^{-1}$. In

other words, with high probability, we have to estimate H given only $1 + \alpha^{-1}$ numbers, namely $\{h_j\}_{0 \leq j \leq \alpha^{-1}}$, and it is not hard to see, given (1.12) and the usual Bayesian lower bounds on minimax error rates (see, e.g., [Ritov and Bickel, 1990]), that this is not enough to estimate $H(p)$. We have, therefore:

Theorem 7. *If $N \sim O(m^{1-\alpha})$, $\alpha > 0$, then no consistent estimator for H exists.*

By Shannon's discrete formula, a similar result holds for mutual information.

1.6 Approximation theory and bias

The last equality — expression (1.12) in the previous section — is key to the rest of our development. Letting $B(\hat{H})$ denote the bias of $\hat{H}$, we have:

$$\begin{aligned} B(\hat{H}) &= (\sum_{j=0}^{N} a_{j,N} \sum_{i=1}^{m} \binom{N}{j} p_i^j (1-p_i)^{N-j}) - (\sum_{i=1}^{m} -p_i \log(p_i)) \\ &= (\sum_i p_i \log(p_i)) + \sum_i \sum_j a_{j,N} \binom{N}{j} p_i^j (1-p_i)^{N-j} \\ &= \sum_i (p_i \log(p_i) + \sum_j a_{j,N} \binom{N}{j} p_i^j (1-p_i)^{N-j}). \end{aligned}$$

If we define the usual entropy function

$$H(x) = -x \log x$$

and the binomial polynomials

$$B_{j,N}(x) \equiv \binom{N}{j} x^j (1-x)^{N-j},$$

we have

$$-B(\hat{H}) = \sum_i (H(p_i) - \sum_j a_{j,N} B_{j,N}(p_i)).$$

In other words, the bias is the m-fold sum of the difference between the function H and a polynomial of degree N; these differences are taken at the points p_i, which all fall on the interval $[0, 1]$. The bias will be small, therefore, if the polynomial is close, in some suitable sense, to H. This type of polynomial approximation problem has been extensively studied [Devore and Lorentz, 1993], and certain results from this general theory of approximation will prove quite useful.

Given any continuous function f on the interval, the Bernstein approximating polynomials of f, $B_N(f)$, are defined as a linear combination of the binomial polynomials defined above:

$$B_N(f)(x) \equiv \sum_{j=0}^{N} f(j/N) B_{j,N}(x).$$

Note that, for the MLE,

$$a_{j,N} = H(j/N);$$

that is, the polynomial appearing in (1.12) is, for the MLE, exactly the Bernstein polynomial for the entropy function $H(x)$. Everything we know about the bias of the MLE (and more) can be derived from a few sim-

ple general facts about Bernstein polynomials. For example, we find the following result in [Devore and Lorentz, 1993]:

Theorem ([Devore and Lorentz, 1993]: 10.4.2). *If f is strictly concave on the interval, then*

$$B_N(f)(x) < B_{N+1}(f)(x) < f(x),\ 0 < x < 1.$$

Clearly, H is strictly concave, and $B_N(H)(x)$ and H are continuous, hence the bias is everywhere nonpositive; moreover, since

$$B_N(H)(0) = H(0) = 0 = H(1) = B_N(H)(1),$$

the bias is strictly negative unless p is degenerate. Of course, we already knew this, but the above result makes the following, less well-known proposition easy:

Proposition 8. *For fixed m and nondegenerate p, the bias of the MLE is strictly decreasing in magnitude as a function of N.*

(Couple the local expansion, equation (1.6), with [Cover and Thomas, 1991], Chapter 2, Problem 34, for a purely information-theoretic proof.)

The second useful result is given in the same chapter:

Theorem ([Devore and Lorentz, 1993]: 10.3.1). *If f is bounded on the interval, differentiable in some neighborhood of x, and has second derivative $f''(x)$ at x, then*

$$\lim_{N\to\infty} N(B_N(f)(x) - f(x)) = f''(x)\frac{x(1-x)}{2}.$$

This theorem hints at the desired generalization of Miller's original result on the asymptotic behavior of the bias of the MLE:

Theorem 9. *If* $m > 1$, $N \min_i p_i \to \infty$, *then*

$$\lim \frac{N}{m-1} B(\hat{H}_{MLE}) = -\frac{1}{2}.$$

The proof is an elaboration of the proof of the above theorem 10.3.1 of [Devore and Lorentz, 1993]; we leave it for the appendix. Note that the convergence stated in the theorem given by [Devore and Lorentz, 1993] is not uniform for $f = H$, because $H(x)$ is not differentiable at $x = 0$; thus, when the condition of the theorem is not met (i.e., $\min_i p_i = O(\frac{1}{N})$), more intricate asymptotic bias formulae are necessary; as before, we can use the Poisson approximation for the bins with $Np_i \to c$, $0 < c < \infty$, and an $o(1/N)$ approximation for those bins with $Np_i \to 0$.

1.6.1 BUB estimator

Theorem 9 suggests one simple way to reduce the bias of the MLE: make the substitution

$$\begin{aligned} a_{j,N} = -\frac{j}{N}\log\frac{j}{N} \quad &\to \quad a_{j,N} - H''(\frac{j}{N})\frac{\frac{j}{N}(1-\frac{j}{N})}{2N} \\ &= \quad -\frac{j}{N}\log\frac{j}{N} + \frac{(1-\frac{j}{N})}{2N}. \end{aligned} \tag{1.13}$$

This leads exactly to a version of the Miller-Madow correction, and gives another angle on why this correction fails in the $N \sim m$ regime: as discussed above, the singularity of $H(x)$ at 0 is the impediment.

A more systematic approach towards reducing the bias would be to choose $a_{j,N}$ such that the resulting polynomial is the best approximant of $H(x)$ within the space of N-degree polynomials. This space corresponds exactly to the class of estimators that are, like $\hat{H}$, linear in the histogram order statistics. We write this correspondence explicitly:

$$\{a_{j,N}\}_{0\leq j\leq N} \longleftrightarrow \hat{H}_{a,N},$$

where we define $\hat{H}_{a,N} \equiv \hat{H}_a$ to be the estimator determined by $a_{j,N}$, according to

$$\hat{H}_{a,n} = \sum_{j=0}^{N} a_{j,N} h_j.$$

Clearly, only a small subset of estimators have this linearity property; the h_j-linear class comprises an $N+1$-dimensional subspace of the m^N-dimensional space of all possible estimators. (Of course, m^N overstates the case quite a bit, as this number ignores various kinds of symmetries we would want to build into our estimator (see Propositions 19 and 20), but it is still clear that the linear estimators do not exhaust the class of all reasonable estimators.) Nevertheless, this class will turn out to be quite useful.

What sense of "best approximation" is right for us? If we are interested in worst-case results, uniform approximation would seem to be a good choice: that is, we want to find the polynomial that minimizes

$$M(\hat{H}_a) \equiv \max_x \ |H(x) - \sum_j a_{j,N} B_{j,N}(x)|.$$

(Note the *the* above: the best approximant in this case turns out to be unique [Devore and Lorentz, 1993], although we will not need this fact below.) A bound on $M(\hat{H}_a)$ obviously leads to a bound on the maximum bias over all p:

$$\max_p |B(\hat{H}_a)| \leq mM(\hat{H}_a).$$

However, the above inequality is not particularly tight; we know, by Markov's inequality, that p can not have too many components greater than $1/m$, and therefore the behavior of the approximant for x near $x = 1$ might be less important than the behavior near $x = 0$. Therefore, it makes sense to solve a weighted uniform approximation problem: minimize

$$M^*(f, \hat{H}_a) \equiv \sup_x \ (f(x)|H(x) - \sum_j a_{j,N} B_{j,N}(x)|),$$

where f is some positive function on the interval. The choice $f(x) = m$, thus, corresponds to a bound of the form

$$\max_p |B(\hat{H}_a)| \leq c^*(f) M^*(f, \hat{H}_a),$$

with the constant $c^*(f)$ equal to one here. Can we generalize this?

According to the discussion above, we would like f to be larger near zero than near one, since p can have many small components but at most $1/x$ components greater than x. One obvious candidate for f, then, is $f(x) = 1/x$: it is easy to prove that

$$\max_p |B(\hat{H}_a)| \leq M^*(1/x, \hat{H}_a),$$

i.e., $c^*(1/x) = 1$ (see appendix). However, this f gives too much weight to small p_i; a better choice is

$$f(x) = \begin{cases} m & x < 1/m, \\ 1/x & x \geq 1/m. \end{cases}$$

For this f, we have

Proposition 10.

$$\max_p |B(\hat{H}_a)| \leq c^*(f) M^*(f, \hat{H}_a)), \ c^*(f) = 2.$$

See the appendix for the proof.

It can be shown, using the above bounds combined with a much deeper result from approximation theory [Devore and Lorentz, 1993, Ditzian and Totik, 1987], that there exists an $a_{j,N}$ such that the maximum (over all p) bias is $O(\frac{m}{N^2})$. This is clearly better than the $O(\frac{m}{N})$ rate offered by the three most popular $\hat{H}$. We even have a fairly efficient algorithm to compute this estimator (a specialized descent algorithm developed by Remes [Watson, 1980]). Unfortunately, the good approximation properties of this estimator are a result of a delicate balancing of large, oscillating coefficients $a_{j,N}$, and the variance of the corresponding estimator turns out to be very large. (This is predictable, in retrospect: we already know that no consistent estimator exists if $m \sim N^{1+\alpha}, \alpha > 0$.) Thus, to find a good estimator, we need to minimize bounds on bias and variance simultaneously;

we would like to find $\hat{H}_a$ to minimize

$$\max_p (B_p(\hat{H}_a)^2 + V_p(\hat{H}_a)),$$

where the notation for bias and variance should be obvious enough. We have

$$\begin{aligned} \max_p (B_p(\hat{H}_a)^2 + V_p(\hat{H}_a)) &\leq \max_p B_p(\hat{H}_a)^2 + \max_p V_p(\hat{H}_a) \\ &\leq (c^*(f)M^*(f, \hat{H}_a))^2 + \max_p V_p(\hat{H}_a), \quad (1.14) \end{aligned}$$

and at least two candidates for easily computable uniform bounds on the variance. The first comes from McDiarmid:

Proposition 11.

$$Var(\hat{H}_a) < N \max_{0 \leq j < N} (a_{j+1} - a_j)^2.$$

This proposition is a trivial generalization of the corresponding result of [Antos and Kontoyiannis, 2001] for the MLE; the proofs are identical. We will make the abbreviation

$$||Da||_\infty^2 \equiv \max_{0 \leq j < N} (a_{j+1} - a_j)^2.$$

The second variance bound comes from Steele; see the appendix for the proof (again, a generalization of the corresponding result for $\hat{H}$):

Proposition 12.

$$Var(\hat{H}_a) < 2c^*(f) \sup_x \left| f(x) \left(\sum_{j=2}^{N} j(a_{j-1} - a_j)^2 B_{j,N}(x) \right) \right|.$$

Thus we have our choice of several rigorous upper bounds on the maximum expected error, over all possible underlying distributions p, of any given $\hat{H}_a$. If we can find a set of $\{a_{j,N}\}$ that makes any of these bounds small, we will have found a good estimator, in the worst-case sense; moreover, we will have uniform conservative confidence intervals with which to gauge the accuracy of our estimates. (Note that Propositions 10, 11, and 12 can be used to compute strictly conservative errorbars for other h_j-linear estimators; all one has to do is plug in the corresponding $\{a_{j,N}\}$.)

Now, how do we find such a good $\{a_{j,N}\}$? For simplicity, we will base our development here on the McDiarmid bound (Proposition 11), but very similar methods can be used to exploit the Steele bound. Our first step is to replace the above L_∞ norms with L_2 norms; recall that

$$M^*(f, H_a)^2 = ||f(H - \sum_j a_{j,N} B_{j,N})||_\infty^2.$$

So to choose $a_{N,j}$ in a computationally feasible amount of time, we minimize the following:

$$c^*(f)^2 ||(f)(H - \sum_j a_{j,N} B_{j,N})||_2^2 + N||Da||_2^2. \tag{1.15}$$

This is a "regularized least-squares" problem, whose closed-form solution is well-known; the hope is that the (unique) minimizer of expression (1.15) is a near-minimizer of expression (1.14), as well. The solution for the best $a_{j,N}$, in vector notation, is

$$a = (X^t X + \frac{N}{c^*(f)^2} D^t D)^{-1} X^t Y, \tag{1.16}$$

where D is the difference operator, defined as in Proposition 11, and X^tX and X^tY denote the usual matrix and vector of self- and cross-products, $< B_{j,N}f, B_{k,N}f >$ and $< B_{j,N}f, Hf >$, respectively.

As is well-known [Press et al., 1992], the computation of the solution (1.16) requires on the order of N^3 time steps. We can improve this to an effectively $O(N)$-time algorithm with an empirical observation: for large enough j, the $a_{N,j}$ computed by the above algorithm look a lot like the $a_{N,j}$ described in expression (1.13) (data not shown). This is unsurprising, given Devore and Lorentz's theorem 10.3.1; the trick we took advantage of in expression (1.13) should work exactly for those j for which the function to be approximated is smooth at $x = \frac{j}{N}$, and $H(\frac{j}{N})$ becomes monotonically smoother as j increases.

Thus, finally, we arrive at an algorithm: for $0 < k < K << N$, set $a_{N,j} = -\frac{j}{N}\log\frac{j}{N} + \frac{(1-\frac{j}{N})}{2N}$ for all $j > k$, and choose $a_{N,j}, j \leq k$ to minimize the least-squares objective function (1.15); this entails a simple modification of (1.16):

$$a_{j\leq k} = \left(X^tX_{j\leq k} + \frac{N}{c^*(f)^2}(D^tD + I_k^tI_k)\right)^{-1}\left(X^tY_{j\leq k} + \frac{Na_{k+1}}{c^*(f)^2}e_k\right),$$

where I_k is the matrix whose entries are all zero, except for a one at (k,k), e_k is the vector whose entries are all zero, except for a one in the k-th element, $X^tX_{j\leq k}$ is the upper-left $k \times k$ submatrix of X^tX, and

$$X^tY_{j\leq k} =< fB_{j,N}, f(H - \sum_{j=k+1}^{N} a_{j,N}B_{j,N}) > .$$

Last, choose $a_{N,j}$ to minimize the true objective function (1.14) over all K estimators so obtained. In practice, the minimal effective K varies quite slowly with N (for example, for $N = m < 10^5$, $K \approx 30$); thus the algorithm is approximately (but not rigorously) $O(N)$. (Of course, once a good $\{a_{j,N}\}$ is chosen, $\hat{H}_a$ is no harder to compute than $\hat{H}$.) We will refer to the resulting estimator as $\hat{H}_{BUB}$, for "best upper bound"; matlab code implementing this estimator is available at http://www.cns.nyu.edu/∼liam.

Before we discuss $\hat{H}_{BUB}$ further, we note several minor but useful modifications of the above algorithm. First, for small enough N, the regularized least-squares solution can be used as the starting point for a hill-climbing procedure, minimizing expression (1.14) directly, for slightly improved results. Second, f, and the corresponding $c^*(f)^{-2}$ prefactor on the variance (D^tD) term, can be modified if the experimenter is more interested in reducing bias than variance, or vice versa. Finally, along the same lines, we can constrain the size of a given coefficient $a_{k,N}$ by adding a Lagrange multiplier to the regularized least-square solution as follows:

$$(X^tX + \frac{N}{c^*(f)^2}D^tD + \lambda_k I_k^t I_k)^{-1}X^tY,$$

where I_k is as defined above. This is useful in the following context: at points p for which $H(p)$ is small, most of the elements of the typical empirical measure are zero; hence the bias near these points is $\approx (N-1)a_{0,N}+a_{N,N}$, and λ_0 can be set as high as necessary to keep the bias as low as desired near these low entropy points. Numerical results show that these perturbations

have little ill effect on the performance of the estimator; for example, the worst-case error is relatively insensitive to the value of λ_0 (see Figure 1.7).

The performance of this new estimator is quite promising. Figure 1.3 indicates that, when m is allowed to grow linearly with N, the upper bound on the RMS error of this estimator (the square root of expression (1.14)) drops off approximately as

$$\max_p((E(\hat{H}_{BUB} - H)^2)^{1/2}) <\sim N^{-\alpha}, \alpha \approx 1/3.$$

(Recall that we have a *lower* bound on the worst-case error of the three most common $\hat{H}$:

$$\max_p((E(\hat{H} - H)^2)^{1/2}) >\sim B_{\hat{H}}(N/m),$$

where $B_{\hat{H}}(N/m)$ is a bias term that remains bounded away from zero if N/m is bounded.) For emphasis, we codify this observation as a conjecture:

Conjecture. *$\hat{H}_{BUB}$ is consistent as $N \to \infty$ even if $N/m \sim c,\ 0 < c < \infty$.*

This conjecture is perhaps not as surprising as it appears at first glance; while, intuitively, the nonparametric estimation of the full distribution p on m bins should require $N >> m$ samples, it is not *a priori* clear that estimating a single parameter, or functional of the distribution, should be so difficult. Unfortunately, while we have been able to sketch a proof of the above conjecture, we have not yet obtained any kind of complete asymptotic theory for this new estimator along the lines of the consistency results of

section (1.4); we hope to return to this question in more depth in the future (see section 1.8.3).

From a nonasymptotic point of view, the new estimator is clearly superior to the three most common $\hat{H}$, even for small N, if N/m is small enough: the upper bounds on the error of the new estimator are smaller than the lower bounds on the worst-case error of $\hat{H}_{JK}$ for $N/m = 1$, for example, by $N \approx 1000$, while the crossover point occurs at $N \approx 50$ for $m = 4N$. (We obtain these lower bounds by computing the error on a certain subset of the parameter space on which exact calculations are possible; see section 1.7. For this range of N and m, $\hat{H}_{JK}$ always had a smaller maximum error than $\hat{H}_{MLE}$ or $\hat{H}_{MM}$.) For larger values of N/m or smaller values of N, the figure is inconclusive, as the upper bounds for the new estimator are greater than the lower bounds for $\hat{H}_{JK}$. However, the numerical results in the next section indicate that in fact $\hat{H}_{BUB}$ performs as well as the three most common $\hat{H}$ even in the $N >> m$ regime.

1.7 Numerical results and applications to data

What is the best way to quantify the performance of this new estimator (and to compare this performance to that of the three most common $\hat{H}$)? Ideally, we would like to examine the expected error of a given estimator simultaneously for all parameter values. Of course, this is only possible when the parameter space is small enough; here, our parameter space is

the $(m-1)$-dimensional space of discrete distributions on m points, so we can directly display the error function only if $m \leq 3$ (Figure 1.4). For larger m, we can either compute upper bounds on the worst-case error, as in the previous section (this worst-case error is often considered the most important measure of an estimator's performance if we know nothing about the *a priori* likelihood of the underlying parameter values), or we can look at the error function on what we hope is a representative slice of the parameter space.

One such slice through parameter space is given by the "central lines" of the m-simplex: these are the subsets formed by linearly interpolating between the trivial (minimal entropy) and flat (maximal entropy) distributions (there are m of these lines, by symmetry). Figure 1.4 shows, for example, that the worst-case error for the MLE is achieved on these lines, and it seems plausible that these lines might form a rich enough class that it is as difficult to estimate entropy on this subset of the simplex as it is on the entire parameter space. While this intuition is not quite correct (it is easy to find reasonable estimators whose maximum error does not fall on these lines), calculating the error on these central lines does at least give us a lower bound on the worst-case error. By recursively exploiting the permutation symmetry and the one-dimensional nature of the problem, we constructed a fast algorithm to exactly compute these central line error functions — explicitly enumerating all possible sorted histograms for a given (m, N) pair via a special recursion, computing the multinomial probability

and estimating the $\hat{H}(p')$ associated with each histogram, and obtaining the desired moments of the error distribution at each point along the central line. The results are shown in Figures 1.5 and 1.6.

Figure 1.5 illustrates two important points. First, the new estimator performs quite well; its maximum error on this set of distributions is about half as large as that of the next best estimator, $\hat{H}_{JK}$, and about a fifth the size of the worst-case error for the MLE. In addition, even in the small region where the error of $\hat{H}$ is less than that of $\hat{H}_{BUB}$ — near the point at which $H = \hat{H} = 0$ — the error of the new estimator remains acceptably small. Second, these exact computations confirm the validity of the bias approximation of Theorem 6, even for small values of N. Compare, for example, the bias predicted by the fixed m, large N theory [Miller, 1955], which is *constant* on the interior of this interval. This figure thus clearly shows that the classical asymptotics break down when the $N >> m$ condition is not satisfied, and that the $N \sim m$ asymptotics introduced in section 1.5 can offer a powerful replacement. Of course, neither approximation is strictly "better" than the other, but one could argue that the $N \sim m$ situation is in fact the more relevant for neuroscientific applications, where m is often allowed to vary with N.

In figure 1.6, we show these central line error curves for a few additional (N, m) combinations. Recall Figure 1.3: if N is too small and N/m is too large, the upper bound on the error of $\hat{H}_{BUB}$ is in fact greater than the lower

bound on the worst-case error for $\hat{H}_{JK}$; thus the analysis presented in the previous section is inconclusive in this (N, m) regime. However, as Figure 1.6 indicates, the new estimator seems to perform well even as N/m becomes large; the maximum error of $\hat{H}_{BUB}$ on the central lines is strictly less than that of the three most common estimators for all observed combinations of N and m, even for $N = 10m$. Remember that all four estimators are basically equivalent as $n_i \to \infty$, where the classical (Cramer-Rao) behavior takes over and variance dominates the mean-square error of the MLE. In short, the performance of the new estimator seems to be even better than the worst-case analysis of section 1.6.1 indicated.

While the central lines are geometrically appealing, they are certainly not the only family of distributions we might like to consider. We examine two more such families in Figure 1.7 and find similar behavior. The first panel shows the bias of the same four estimators along the flat distributions on m' bins, $1 \leq m' \leq m$, where as usual only m and N are known to the estimator. Note the emergence of the expected log-linear behavior of the bias of $\hat{H}$ as N/m becomes small (recall the discussion following Theorem 6). The second panel shows the bias along the family $p_i \simeq i^{\alpha}$, for $0 < \alpha < 20$, where similar behavior is evident. This figure also illustrates the effect of varying the λ_0 parameter: the bias at low entropy points can be reduced to arbitrarily low levels at the cost of relatively small changes in the bias at the high-entropy points on the m-simplex. As above, the Steele bounds on the

variance of each of these estimators was comparable, with $\hat{H}_{BUB}$ making a modest sacrifice in variance to achieve the smaller bias shown here.

One could object that the set of probability measures examined in Figures 1.5, 1.6, and 1.7 might not be relevant for neural data; it is possible, for example, that probability measures corresponding to cellular activity lie in a completely different part of parameter space. In the next figure, therefore, we examined our estimators' behavior over a range of p generated by the most commonly used neural model, the integrate-and-fire (IF) cell. The exact calculations presented in the previous figures are not available in this context, so we turned to a Monte Carlo approach. We drove an IF cell with i.i.d. samples of Gaussian white noise, discretized the resulting spike trains in binary fashion (with discretization parameters comparable to those found in the literature), and applied the four estimators to the resulting binned spike trains.

Figure 1.8 shows the bias, variance, and RMS error of our four estimators over a range of parameter settings, in a spirit similar to that of Figure 1.5; the critical parameter here was the mean firing rate, which was adjusted by systematically varying the DC value of the current driving the cell. (Because we are using simulated data, we can obtain the "true" value of the entropy simply by increasing N until $\hat{H}$ is guaranteed to be as close as desired to the true H, with probability approaching one.) Note that as the DC current increases, the temporal properties of the spike trains

changes as well; at low DC, the cells are essentially noise-driven, and have a correspondingly randomized spike train (as measured, e.g., by the coefficient of variation of the inter-spike interval distribution), while at high DC the cells fire essentially periodically (low ISI coefficient of variation). The results here are similar to those in the previous two figures: the bias of the new estimator is drastically smaller than that of the other three estimators over a large region of parameter space. Again, when $H(p) \to 0$ (this occurs in the limit of high firing rates — when all bins contain at least one spike — and low firing rates, where all bins are empty), the common estimators outperform $\hat{H}_{BUB}$, but even here, the new estimator has acceptably small error.

Finally, we applied our estimators to two sets of real data (Figs. 1.9 and 1.10). The *in vitro* data set (Fig. 1.9) was recorded in the lab of Alex Reyes: in a rat cortical slice preparation, we obtained double whole-cell patches from single cells. We injected a white-noise current stimulus via one electrode while recording the voltage response through the other electrode. Recording and data processing followed standard procedures; see [Paninski et al., 2003c] for more detail. The resulting spike trains were binned according to the parameters given in the figure legend, which were chosen, roughly, to match values which have appeared in the literature. Results shown are from multiple experiments on a single cell; the standard deviation of the current noise was varied from experiment to experiment to explore different input ranges and firing rates. The *in vivo* dataset

(Fig. 1.10) was recorded in the lab of John Donoghue: we recorded simultaneously from multiple cells in the arm representation of the primary motor cortex while a monkey moved its hand according to a stationary, two-dimensional, filtered Gaussian noise process. We show results for 11 cells, simultaneously recorded during a single experiment (Fig. 1.10); note, however, that we are estimating the entropy of single-cell spike trains, not the full multi-cell spike train. See [Paninski et al., 1999, Paninski, 2003b] for more details on the experimental procedures.

With real data, it is of course impossible to determine the true value of H, and so the detailed error calculations performed above are not possible here. Nevertheless, the behavior of these estimators seems to follow the trends seen in the simulated data: we see the consistent slow increase in our estimate as we move from $\hat{H}_{MLE}$ to $\hat{H}_{MM}$ to $\hat{H}_{JK}$, and then a larger jump as we move to $\hat{H}_{BUB}$. This is true even though the relevant time scales (roughly defined as the correlation time of the stimulus) in the two experiments differed by about three orders of magnitude. Similar results were obtained for both the real and simulated data using a variety of other discretization parameters (data not shown). Thus, as far as can be determined, our conclusions about the behavior of these four estimators, obtained via the analytical and numerical techniques described above, seem to be consistent with results obtained using physiological data.

In all, we have that the new estimator performs quite well in a uniform sense. This good performance is especially striking, but not limited to,

the case when N/m is $O(1)$. We emphasize that even at the points where $H = \hat{H} = 0$ (and therefore the three most common estimators perform well, in a trivial sense), the new estimator performs reasonably; by construction, $\hat{H}_{BUB}$ never exhibits blowups in the expected error like those seen with $\hat{H}_{MLE}$, $\hat{H}_{MM}$, and $\hat{H}_{JK}$. Given the fact that we can easily tune the bias of the new estimator at points where $H \approx 0$, by adjusting λ_0, $\hat{H}_{BUB}$ appears to be a robust and useful new estimator. We offer matlab code, at http://www.cns.nyu.edu/∼liam, to compute the exact bias and Steele variance bound for any $\hat{H}_a$, at any distribution p, if the reader is interested in more detailed investigation of the properties of this class of estimator.

1.8 Directions for future work

We have left a few important open problems. Below, we give three somewhat freely defined directions for future work, along with a few preliminary results.

1.8.1 Bayes

All of our results here have been from a minimax, or "worst-case," point of view. As discussed above, this approach is natural if we know very little about the underlying probability measure. However, in many cases, we do know something about this underlying p - we might know that the spike count is distributed according to something like a Poisson distribution, or

that the responses of a neuron to a given set of stimuli can be fairly well approximated by a simple dynamical model, such as an integrate-and-fire cell. How do we incorporate this kind of information in our estimates? The fields of parametric and Bayesian statistics address this issue explicitly. We have not systematically explored the parametric point of view — this would entail building a serious parametric model for spike trains and then efficiently estimating the entropy at each point in the parameter space — although this approach has been shown to be powerful in a few select cases. The Bayesian approach would involve choosing a suitable *a priori* distribution on spike trains and then computing the corresponding MAP or conditional mean estimator; this approach is obviously difficult as well, and we can only give a preliminary result here.

Wolpert and Wolf [Wolpert and Wolf, 1995] give an explicit formula for the Bayes' estimate of H and related statistics in the case of a uniform prior on the simplex. We note an interesting phenomenon relevant to this estimator: as m increases, the distribution on H induced by the flat measure on the simplex becomes concentrated around a single point, and therefore the corresponding Bayes' problem becomes trivial as $m \to \infty$, quite the opposite of the situation considered in the current work. ([Nemenman et al., 2002] independently obtained a few interesting results along these lines.) The result is interesting in its own right; its proof shares many of the features (concentration of measure and symmetry techniques) of our main results in the preceding sections. More precisely:

We consider a class of priors determined by the following "sort-difference" procedure: fix some probability measure P on the unit interval. Choose $m-1$ independent samples distributed according to P; sort the samples in ascending order, and call the sorted samples $\{x_i\}_{0<i<m}$. Define $q_1 = x_1$, $q_m = 1 - x_{m-1}$, and $q_i = x_i - x_{i-1}$ for all other i. This procedure therefore generates random probability measures q on m bins; in different language, the sort-difference procedure induces a prior on the m-simplex. (If P is the uniform density on the interval, for example, this prior is uniform on the m-simplex; this is the main case considered in [Wolpert and Wolf, 1995].) The prior on q induces a prior on H, and this prior on H, in turn, happens to have a surprisingly small variance, for reasons quite similar to the reasons $\hat{H}$ has a surprisingly small variance: the entropy functional $H(p)$ is a symmetric and fairly smooth functional of p. So, let the prior on H, $P(H)$, be generated by this sort-difference procedure, and assume for technical simplicity that the interval measure $P[0,1]$ has a density component, p. We have the following crude but interesting result:

Theorem 13. *If p is bounded away from zero, then H is normally concentrated with rate $m^{1/3}$; that is, for fixed a,*

$$p(|H - E(H)| > a) = O(e^{-Cm^{1/3}a^2}),$$

for any constant $a > 0$ and some constant C.

In fact, it is possible to prove much more: the uniform measure on the

simplex (and more generally, any prior induced by the sort-difference procedure, under some conditions on the interval measure P) turns out to induce an asymptotically normal prior on H, with variance decreasing in m. We can calculate the asymptotic mean of this distribution by using linearity of expectation and symmetry techniques like those used in section 1.5. In the following, assume for simplicity that P is equivalent to Lebesgue measure (that is, P is absolutely continuous with respect to Lebesgue measure, and vice versa); this is a technical condition which can be relaxed at the price of slightly more complicated formulae. We have the following:

Theorem 14. *P(H) is asymptotically normal, with*

$$Var(H) \sim \frac{1}{m}$$

and asymptotic mean calculated as follows.

Let q be the sorted, normalized density corresponding to a measure drawn according to the prior described above; define

$$F_p(v) \equiv \int_0^v du \int_0^1 dt p(t)^2 e^{-up(t)},$$

and

$$q'_\infty \equiv F_p^{-1},$$

where the inverse is taken in a distributional sense. Then

$$||q - q'_\infty||_1 \to 0$$

in probability and

$$E(H) \to h(q'_\infty) + \log(m),$$

where $h(.)$ denotes differential entropy.

F_p above is the cumulative distribution function of the p-mixture of exponentials with rate $p(t)^{-1}$ (just as $p'_{c,\infty}$ in Theorem 6 was defined as the inverse c.d.f. of a mixture of Poisson distributions). If P is uniform, for example, we have that $||q - q'_\infty||_1 \to 0$ in probability, where

$$q'_\infty(t) = -\log(t),$$

and

$$H(q) \to h(q'_\infty) + \log(m) = \log m + \int_0^1 dt \log(t) \log(-\log(t))$$

in probability.

1.8.2 Adaptive partitioning

As emphasized in the introduction, we have restricted our attention here to partitions, "sieves," S and T, which do not depend on the data. This is obviously a strong condition. Can we obtain any results without this assumption?

As a start, we have the following consistency result, stated in terms of the measure of the richness of a partition introduced by Vapnik and Chervonenkis, $\Delta_N(\mathcal{A}_\mathcal{F})$ (the shatter coefficient of the set of allowed partitions, defined in the appendix; m is, as in the preceding, the maximal number of elements per partition, $\mathcal{F}$; [Devroye et al., 1996]):

Theorem 15. *If* $\log \Delta_N(\mathcal{A}_\mathcal{F}) = \mathrm{o}(\frac{\mathrm{N}}{(\log \mathrm{m})^2})$ *and* $\mathcal{F}$ *generates* $\sigma_{x,y}$ *a.s.,* $\hat{I}$ *is consistent in probability;* $\hat{I}$ *is consistent a.s. under the slightly stronger condition*

$$\sum \Delta_N(\mathcal{A}_\mathcal{F}) \mathrm{e}^{\frac{-\mathrm{N}}{(\log \mathrm{m})^2}} < \infty.$$

Note the slower allowed rate of growth of m. In addition, the conditions of this theorem are typically harder to check than those of Theorem 2. For example, it is easy to think of reasonable partitioning schemes which do not generate $\sigma_{x,y}$ a.s.: if the support of P is some measurable proper subset of $X \times Y$, this is an unreasonable condition. We can avoid this problem by rephrasing the condition in terms of $\sigma_{x,y}$ restricted to the support of P (this, in turn, requires placing some kind of topology on $X \times Y$, which should be natural enough in most problems).

What are the benefits? Intuitively, we should gain in efficiency: we are putting the partitions where they do the most good [Darbellay and Vajda, 1999]. We also gain in applicability, since in practice all partition schemes are data-driven to some degree. The most important application of this result, however, is to the following question in learning theory: how do we choose the most informative partition? For example, given a spike train and some behaviorally relevant signal, what is the most efficient way to encode the information in the spike train about the stimulus? More concretely: all things being equal, does encoding temporal information, say, preserve more information about a given visual stimulus than

encoding spike rate information? Conversely, does encoding the contrast of a scene, for example, preserve more information about a given neuron's activity than encoding color? Given m codewords, how much information can we capture about what this neuron is telling us about the scene? See, e.g., [Victor, 2000b] for recent work along these lines.

The formal analog to these kinds of questions is as follows (see [Tishby et al., 1999] and [Gedeon et al., 2003] for slightly more general formulations). Let $\mathcal{F}$ and $\mathcal{G}$ be classes of "allowed" functions on the spaces $\mathcal{X}$ and $\mathcal{Y}$. For example, $\mathcal{F}$ and $\mathcal{G}$ could be classes of partitioning operators (corresponding to the discrete setup used here), or spaces of linear projections (corresponding to the information-maximization approach to ICA). Then, given N i.i.d. data pairs in $X \times Y$, we are trying to choose $f_N \in \mathcal{F}$ and $g_N \in \mathcal{G}$ in such a way that

$$I(f_N(x); g_N(y))$$

is maximized. This is where results like Theorem 15 are useful; they allow us to place distribution-free bounds on

$$P(\sup_{f \in \mathcal{F}, g \in \mathcal{G}} I(f(x); g(y)) - \hat{I}(f_N(x); g_N(y)) > \epsilon), \tag{1.17}$$

i.e., the probability that the set of codewords that looks optimal given N samples is actually ϵ-close to optimal. Other (distribution-dependent) approaches to the asymptotics of quantities like (1.17) come from the the theory of empirical processes; see, e.g., [van der Vaart and Wellner, 1996].

More work in this direction will be necessary to rigorously answer the bias and variance problems associated with these "optimal coding" questions.

1.8.3 Smoothness and other functionals

We end with a slightly more abstract question. In the context of the sieve method analyzed here, are entropy and mutual information any harder to estimate than any other given functional of the probability distribution? Clearly, there is nothing special about H (and by extension I) in the case when m and p are fixed; here, classical methods lead to the usual $N^{-1/2}$ rates of convergence, with a prefactor that only depends on m and the differential properties of the functional H at p; the entire basic theory goes through if H is replaced by some other arbitrary (smooth) functional.

There are several reasons to suspect, however, that not all functionals are the same when m is allowed to vary with N. First, most obviously, $\hat{H}$ is consistent when $m = o(N)$ but not when $m \sim N$; simple examples show that this is not true for all functionals of p (for example, many linear functionals on m can be estimated given fewer than N samples, and this can be extended to weakly nonlinear functionals as well). Second, classical results from approximation theory indicate that smoothness plays an essential role in approximability; it is well-known, for example, that the best rate in the bias polynomial approximation problem described in section 1.6 is essentially determined by a modulus of continuity of the function under question [Ditzian and Totik, 1987], and moduli of continu-

ity pop up in apparently very different functional estimation contexts as well [Donoho and Liu, 1991, Jongbloed, 2000]. Thus, it is reasonable to expect that the smoothness of H, especially as measured at the singular point near 0, should have a lot to do with the difficulty of the information estimation problem. Finally, basic results in learning theory [Devroye et al., 1996, van der Vaart and Wellner, 1996, Cucker and Smale, 2002] emphasize the strong connections between smoothness and various notions of learnability. For example, an application of Theorem II.2.3 of [Cucker and Smale, 2002] gives the exact rate of decay of our L_2 objective function (1.15) in terms of the spectral properties of the discrete differential operator D, expressed in the Bernstein polynomial basis; however, it is unclear at present whether this result can be extended to our final goal of a useful asymptotic theory for the upper L_∞ bound (1.14).

A few of the questions we would like to answer more precisely are as follows. First, we would like to have the precise minimax rate of the information estimation problem; thus far, we have only been able to bound this rate between $m \sim o(N)$ (Theorem 2) and $m \sim N^{1+\alpha}$, $\alpha > 0$ (Theorem 7). Second: how close does $\hat{H}_{BUB}$ come to this minimax rate; indeed, does this estimator require fewer than m samples to learn the entropy on m bins, as Figure 1.3 seems to indicate? Finally, how can all of this be generalized for other statistical functionals? Is there something like a single modulus of continuity that controls the difficulty of some large class of these functional estimation problems?

1.9 Conclusions

Several practical conclusions follow from the results presented here; we have good news and bad news. First, the bad news.

- Past work in which N/m was of order 1 or smaller was most likely contaminated by bias, *even if the jackknife or Miller correction was used*. This is particularly relevant for studies in which multiple binning schemes were compared to investigate, e.g., the role of temporal information in the neural code. We emphasize for future studies that m and N *must* be provided for the reader to have confidence in the results of entropy estimation.

- Error bars based on sample variance (or resampling techniques) give very bad confidence intervals if m and N are large; that is, confidence intervals based on the usual techniques *do not* contain the true value of H or I with high probability. Previous work in the literature often displays error bars that are probably misleadingly small. Confidence intervals should be of size

$$\sim B(\hat{H}, N/m) + N^{-1/2} \log(\min(m, N)),$$

  where the bias term B can be calculated via techniques described in section 1.5.

Now the good news:

- This work has given us a much better understanding of exactly how difficult the information estimation problem is, and what we can hope to accomplish using nonparametric techniques, given physiologically plausible sample sizes.

- We have obtained rigorous (and surprisingly general) results on bias, variance, and convergence of the most commonly employed estimators, including the best-possible generalization of Miller's well-known $\frac{1}{N}$ bias rate result. Our analysis clarifies the relative importance of minimizing bias or variability depending on N and m, according to the bias-variance balance function introduced at the end of section 1.4.

- We have introduced a promising new estimator, one which comes equipped with built-in, rigorous confidence intervals. The techniques used to derive this estimator also lead to rigorous confidence intervals for a large class of other estimators (including the three most common $\hat{H}$).

Appendix A: Additional results

A.1 Support

One would like to build an estimator which takes values strictly in some nice set around the true H, say an interval containing H whose length

shrinks as the number of samples, N, increases. This would give us strong "error bars" on our estimate of H - we would be absolutely certain that our estimate is close to the true H. The MLE for entropy has support on $[0, \log(\min(N, m))]$. A simple variational argument shows that any estimator, T, for H on m bins is inadmissible if T takes values outside of $[0, \log m]$. Similarly, any estimator, T, for I on $m_S \times m_T$ bins is inadmissible if T takes values outside of $[0, \log(\min(m_S, m_T))]$. It turns out that this is the best possible, in a sense: there do not exist any nontrivial estimators for entropy which are strictly greater or less than the unknown H. In fact, the following is true:

Proposition 16. *There is no estimator T and corresponding a, b, $0 < a <= 1$, $1 <= b < \infty$, such that the support of T is the interval $[aH, bH]$, for all values of the entropy, H.*

Proof. Suppose such an $a > 0$ exists. If so, $T(\omega)$ must be nonzero for all possible values of the data, ω (the data can be represented as an N-sequence of integers, $1 \leq \omega(i) \leq m$). But then there must exist some ω_0, with $p(\omega_0) > 0$, such that $T(\omega_0) > 0$. By choosing H such that $0 < H < T(\omega_0)$, we force a contradiction. The proof for b is similar. □

A similar result obviously holds for mutual information.

A.2 Bias

It turns out that no unbiased estimators for H or I exist in the discrete setting. This fact seems to be known among information theorists, but we have not seen it stated in the literature. The proof is quite short, so we provide it here.

Proposition 17. *No unbiased estimator for entropy or mutual information exists.*

Proof. For any estimator T of the entropy of a multinomial distribution, we can write down the mean of T:

$$E(T) = \sum_{\omega \in \{1,...,m\}^N} P(\omega)T(\omega),$$

where $\{1, ..., m\}^N$ is the sample space (i.e, each ω, as above, corresponds to an m-ary sequence of length N). Since ω_j is drawn i.i.d. from the discrete distribution p, $P(\omega)$ is given by

$$P(\omega) = \prod_{j=1}^{N} p_{\omega_j},$$

and so the mean of T is a polynomial function of the multinomial probabilities p_i. The entropy, on the other hand, is obviously a nonpolynomial function of the p_i. Hence no unbiased estimator exists. The proof for I is identical. □

The next (easy) proposition provides some more detail. The proof is similar to that of Proposition 16 and is therefore omitted.

Proposition 18. *a. If T is a nonnegatively biased estimator for the entropy of a multinomial distribution on m bins, with $T(\omega) \in [0, \log(m)]\ \forall\ \omega \in \{1, ..., m\}^N$, then*

$$T(\omega) = \log(m)\ \forall\ \omega \in \{1, ..., m\}^N.$$

b. If T is a nonpositively biased estimator for the mutual information of a multinomial distribution on m_S, m_T bins, with $T(\omega) \in [0, \log(\min(m_S, m_T))]$, then

$$T(\omega) = 0\ \forall\ \omega \in \Omega.$$

c. If T is a nonnegatively biased estimator for the mutual information of a multinomial distribution on m_S, m_T bins, with $T(\omega) \in [0, \log(\min(m_S, m_T))]$, then

$$T(\omega) = \log(\min(m_S, m_T))\ \forall\ \omega \in \Omega.$$

A.3 Minimax properties of σ-symmetric estimators

Let the error metric $D(T, \theta)$ be nice — convex in T, jointly continuous in T and θ, positive away from $T = \theta$, and bounded below. (The metrics given by

$$D(T, \theta) \equiv (T - \theta)^p,\ 1 \leq p < \infty$$

are good examples.) The following result partially justifies our focus throughout this paper on estimators which are permutation-symmetric (denoted σ-symmetric in the following).

Proposition 19. *If the error metric D is nice, then a σ-symmetric minimax estimator exists.*

Proof. Existence of a minimax estimator (see also [Schervish, 1995]): whenever $\max_\theta E_\theta(D)$ is a continuous function of the estimator T, a minimax estimator exists, since T can be taken to vary over a compact space (namely, $[0, \log m]^{m^N}$). But $\max \theta(E_\theta(D))$ is continuous in T whenever $E(D)$ is jointly continuous in T and θ. This is because $E(D)$ is uniformly continuous in θ and T, since, again, θ and T vary over compact spaces. $E(D)$ is jointly continuous in θ and T by the continuity of D and the fact that $E(D)$ is defined by a finite sum.

Existence of symmetric minimax estimator: this is actually a special case of the Hunt-Stein theorem [Schervish, 1995]. Any asymmetric minimax estimator, T, in the current setup achieves its maximum, $\max_\theta(E_\theta(D))$, by the arguments above. However, the corresponding symmetrized estimator, $T_\sigma(\omega) = (1/|\sigma|) \sum_\sigma T(\sigma(\omega))$ has expected error which is less than or equal to $\max_\theta(E_\theta(D))$, as can be seen after a rearrangement and an application of Jensen's inequality. Therefore, T_σ is minimax (and obviously symmetric). □

A.4 Insufficiency of symmetric estimators

The next result is perhaps surprising.

Proposition 20. *The MLE is not sufficient. In fact, the empirical his-*

tograms are minimal sufficient; thus, no σ-symmetric estimator is sufficient.

Proof. A simple example suffices to prove the first statement. Choose as a prior on p:

$$P(p(1) = \epsilon; p(2) = 1 - \epsilon) = .5$$
$$P(p(1) = 0; p(2) = 1) = .5,$$

for some $\epsilon > 0$. For this P, $H(p) \to \hat{H} \to \{n_i\}$ does not form a Markov chain; the symmetry of $\hat{H}$ discards information about the true underlying H (namely, observation of a 1 tells us something very different than observation of a 2). This property is clearly shared by any symmetric estimator.

The fact that the empirical histograms are minimal sufficient follows, e.g., from Bahadur's Theorem [Schervish, 1995] and the fact that the empirical histograms are complete sufficient statistics. □

In other words, any σ-symmetric estimator necessarily discards information about H, even though H itself is σ-symmetric. This indicate the importance of priors; the nonparametric minimax approach taken here (focusing strictly on symmetric estimators for a large part of the work, as justified by Proposition 19 above) should only be considered a first step. To be more concrete, in many applications it is natural to guess that the underlying measure p has some continuity properties; therefore, estimators which take advantage of some underlying notion of continuity (for example,

by locally smoothing the observed distributions before estimating their entropy) should be expected to perform better (on average, according to this mostly-continuous prior) than the best σ-symmetric estimator, which necessarily discards all topological structure in the underlying space $\mathcal{X}$. See, e.g., [Victor, 2002] for recent work along these lines.

Appendix B: Proofs

We collect some deferred proofs here; to conserve space, we will omit some of the more easy-to-verify details. The theorems are restated for convenience.

B.1 Consistency

Statement (Theorem 2). *If* $m_{S,N} m_{T,N} = o(N)$ *and* σ_{S_N,T_N} *generates* $\sigma_{X,Y}$*, then* $\hat{I} \to I$ *a.s. as* $N \to \infty$.

Theorem 2 is a consequence of the following lemma:

Statement (Lemma 4). *If* $m = o(N)$*, then* $\hat{H} \to H_N$ *a.s.*

Proof. First, by the exponential bound of Antos and Kontoyiannis (expression (1.4)) and the Borel-Cantelli lemma, $\hat{H}_N \to H_N$ *a.s.* if the (non-random) function $E(\hat{H}_N) \uparrow H_N$. This convergence in expectation is a consequence of the local expansion for the bias of the MLE (expression (1.7)) and proposition 1 of section 1.4. □

Proof of Theorem 2. First, some terminology: by $\hat{I} \to I$ *a.s.*, we mean that, if $I = \infty$, $p((\hat{I}_N < c)\ i.o.) = 0\ \forall c < \infty$, and if $I < \infty$, $p((|\hat{I}_N - I| > \epsilon)\ i.o.) = 0\ \forall \epsilon > 0$. ("I.o." stands for "infinitely often.") In addition, we call the given σ-algebra on $X \times Y$ (the family of sets on which the probability measure $P(X,Y)$ is defined) $\sigma_{X,Y}$, and the sub-σ-algebra generated by S and T $\sigma_{S,T}$.

Now, the proof: it follows from Shannon's formula for mutual information in the discrete case that

$$|\hat{I}(S_N,T_N) - I(S_N,T_N)| \leq |\hat{H}(S) - H(S)| + |\hat{H}(T) - H(T)| + |\hat{H}(S,T) - H(S,T)|.$$

Thus, the lemma gives

$$\hat{I}_N \to I(S_N,T_N)\ a.s.$$

whenever $m_S m_T / N \to 0$.

It only remains to show that the (non-random) function $I(S_N,T_N) \to I$; this follows from results in standard references such as [Billingsley, 1965, Kolmogorov, 1993], if either:

$$\sigma_{S_1,T_1} \subseteq \sigma_{S_2,T_2} \subseteq ... \subseteq \sigma_{S_N,T_N} \subseteq ...$$

and

$$\sigma_{X,Y} = \cup_N \sigma_{S_N,T_N},$$

or

$$\sup_{A \in \sigma_{S_N,T_N}, B \in \sigma_{X,Y}} \rho(A,B) \to 0,$$

where

$$\rho(A, B) \equiv P(AB^c \cup A^c B).$$

If either of these conditions holds, we say that σ_{S_N,T_N} generates $\sigma_{X,Y}$. □

B.2 CLT

Statement (Theorem 3). *Let*

$$\sigma_N^2 \equiv Var(-\log p_{T_N}) \equiv \sum_{i=1}^{m} p_{T_N,i}(-\log p_{T_N,i} - H_N)^2.$$

If $m_N \equiv m = o(N^{1/2})$, *and*

$$\liminf_{N\to\infty} N^{1-\alpha}\sigma_N^2 > 0$$

for some $\alpha > 0$, *then* $(\frac{N}{\sigma_N^2})^{1/2}(\hat{H} - H_N)$ *is asymptotically standard normal.*

Proof. The basic tool, again, is the local expansion of H_{MLE}, expression (1.6). We must first show that the remainder term becomes negligible in probability on a $\sqrt{N}$ scale, that is,

$$\sqrt{N} D_{KL}(p_N; p) = o_p(1).$$

This follows from the formula for $E_p(D_{KL}(p_N; p))$, then Markov's inequality and the nonnegativity of D_{KL}.

So it only remains to show that $dH(p; p_N - p)$ is asymptotically normal. Here we apply a classical theorem on the asymptotic normality of double arrays of infinitesimal random variables:

Lemma. *Let $\{x_{j,N}\}, 1 \leq N \leq \infty, 1 \leq j \leq N$ be a double array of rowwise i.i.d. random variables with zero mean and variance σ_N^2, with distribution $p(x, N)$ and satisfying $\sigma_N^2 = 1/N$ for all N. Then $\sum_{j=1}^{N} x_{j,N}$ is asymptotically normal, with zero mean and unit variance, iff $\{x_{j,N}\}$ satisfy the Lindeberg (vanishing tail) condition: for all $\epsilon > 0$,*

$$\sum_{j=1}^{N} \int_{|x|>\epsilon} x^2 dp(x, N) = o(1). \tag{1.18}$$

The conditions of the theorem imply the Lindeberg condition, with $\{x_{j,n}\}$ replaced by $\frac{1}{\sqrt{N\sigma^2}}(dH(p; \delta_j - p) - H)$. To see this, note that the left hand side of equation (1.18) becomes, after the proper substitutions,

$$\frac{1}{\sigma^2} \sum_{p_j:(N\sigma^2)^{-\frac{1}{2}}|(\log p_j)-H|>\epsilon} p_j \log^2 p_j,$$

or

$$\frac{1}{\sigma^2}\Big(\sum_{p_j:p_j>e^{\epsilon(N\sigma^2)^{\frac{1}{2}}+H}} p_j \log^2 p_j + \sum_{p_j:p_j<e^{H-\epsilon(N\sigma^2)^{\frac{1}{2}}}} p_j \log^2 p_j\Big).$$

The number of terms in the sum on the left is less than or equal to

$$e^{-\epsilon(N\sigma^2)^{\frac{1}{2}}-H};$$

since the summands are bounded uniformly, this sum is $o(1)$. On the other hand, the sum on the right has at most m terms, so under the conditions of the theorem, this term must go to zero as well, and the proof is complete.

□

We have proven the above a.s. and $\sqrt{N}$ consistency theorems for $\hat{H}_{MLE}$ only; the extensions to $\hat{H}_{MM}$ and $\hat{H}_{JK}$ are easy and are therefore omitted.

B.3 Variance bounds a lá Steele

For the σ-symmetric statistic $H_a(\{x_j\})) = \sum_j a_{j,N} h_{j,N}$, Steele's inequality reads:

$$Var(H_a) \leq \frac{N}{2} E((H_a(\{x_j\}) - H_a(x_1, \ldots, x_{N-1}, x'_N))^2),$$

where x'_N is a sample drawn independently from the same distribution as x_j. The linear form of H_a allows us to exactly compute the right hand side of the above inequality. We condition on a given histogram, $\{n_i\}_{i=1,\ldots,m}$:

$$\begin{aligned} E\Big((H_a(\{x_j\}) &- H_a(x_1, \ldots, x_{N-1}, x'_N))^2 \Big) \\ &= \sum_{\{n_i\}} p(\{n_i\}) E\Big((H_a(\{x_j\}) - H_a(x_1, \ldots, x_{N-1}, x'_N))^2 | \{n_i\} \Big). \end{aligned}$$

Now we rewrite the inner expectation on the right-hand side:

$$E((H_a(\{x_j\}) - H_a(x_1, \ldots, x_{N-1}, x'_N))^2 | \{n_i\}) = E((D_- + D_+)^2 | \{n_i\}),$$

where

$$D_- \equiv a_{n_{x_N}-1,N} - a_{n_{x_N},N}$$

is the change in $\sum_j a_{j,N} h_{j,N}$ that occurs when a random sample is removed from the histogram $\{n_i\}$, according to the probability distribution $\{n_i\}/N$, and D_+ is the change in $\sum_j a_{j,N} h_{j,N}$ that occurs when a sample is randomly (and conditionally independently, given $\{n_i\}$) added back to the x_N-less histogram $\{n_i\}$, according to the true underlying measure p_i.

The necessary expectations are as follows. For $1 \leq j \leq N$, define

$$D_j \equiv a_{j-1} - a_j.$$

Then

$$E(D_-^2|\{n_i\}) = \sum_i \frac{n_i}{N} D_{n_i}^2,$$

$$E(D_+^2|\{n_i\}) = \sum_i p_i(\frac{n_i}{N} D_{n_i}^2 + (1 - \frac{n_i}{N}) D_{n_i+1}^2),$$

and

$$\begin{aligned} E(D_+D_-|\{n_i\}) &= E(D_+|\{n_i\})E(D_-|\{n_i\}) \\ &= -(\sum_i \frac{n_i}{N} D_{n_i})(\sum_i p_i(\frac{n_i}{N} D_{n_i} + (1 - \frac{n_i}{N}) D_{n_i+1})). \end{aligned}$$

Taking expectations with respect to the multinomial measure $p(\{n_i\})$, we have

$$E(D_-^2) = \sum_{i,j} \frac{j}{N} D_j^2 B_j(p_i), \tag{1.19}$$

$$E(D_+^2) = \sum_{i,j} (\frac{j}{N} D_j^2 + (1 - \frac{j}{N}) D_{j+1}^2) p_i B_j(p_i),$$

and

$$E(D_+D_-) = -\sum_{i,i',j,k} \frac{j}{N} D_j (\frac{k}{N} D_k + (1 - \frac{k}{N}) D_{k+1}) p_{i'} B_{j,k}(p_i, p_{i'}),$$

where B_j and $B_{j,k}$ denote the binomial and trinomial polynomials, respectively:

$$B_j(t) \equiv \binom{N}{j} t^j (1-t)^{N-j};$$

$$B_{j,k}(s,t) \equiv \binom{N}{j,k} s^j t^k (1-s-t)^{N-j-k}.$$

The obtained bound,

$$Var(H_a) \leq \frac{N}{2} \bigg(E(D_-^2) + 2E(D_-D_+) + E(D_+^2) \bigg),$$

may be computed in $O(N^2)$ time. For a more easily computable ($O(N)$) bound, note that $E(D_-^2) = E(D_+^2)$ and apply Cauchy-Schwartz to obtain

$$Var(H_a) \leq 2NE(D_-^2).$$

Under the conditions of Theorem 3, this simpler bound is asymptotically tight to within a factor of two. Proposition 12 is proven with devices identical to those used in obtaining the bias bounds of Proposition 10 (note the similarity of equations 1.12 and 1.19).

B.4 Convergence of sorted empirical measures

Statement (Theorem 6). *Let P be absolutely continuous with respect to Lebesgue measure on the interval $[0,1]$, and let $p = dP/dm$ be the corresponding density. Let S_N be the m-equipartition of $[0,1]$, p' denote the sorted empirical measure, and $N/m \to c,\ 0 < c < \infty$. Then:*

a) $p' \overset{L_1,a.s.}{\to} p'_{c,\infty}$, with $||p'_{c,\infty} - p||_1 > 0$. Here $p'_{c,\infty}$ is the monotonically decreasing step density with gaps between steps j and $j+1$ given by

$$\int_0^1 dt e^{-cp(t)} \frac{(cp(t))^j}{j!}.$$

b) Assume p is bounded. Then $\hat{H} - H_N \to B_{c,\hat{H}}(p)$ a.s., where $B_{c,\hat{H}}(p)$ is a deterministic function, nonconstant in p. For $\hat{H} = \hat{H}_{MLE}$,

$$B_{c,\hat{H}}(p) = h(p') - h(p) < 0,$$

where $h(.)$ denotes differential entropy.

Proof. We'll give only an outline. By $\overset{L_1,a.s.}{\to}$, we mean that $||p-p_N||_1 \to 0$ $a.s.$ By McDiarmid's inequality, for any distribution q,

$$||q - p_N||_1 \to E(||q - p_N||_1) \ a.s.,$$

Therefore, convergence to some p' in probability in L_1 implies almost sure convergence in L_1. In addition, McDiarmid's inequality hints at the limiting form of the ordered histograms: we have

$$sort(\frac{n_i}{N}) \to E(sort(\frac{n_i}{N})), \ a.s., \ \forall 1 \leq i \leq m.$$

Of course, this isn't quite satisfactory, since both $sort(\frac{n_i}{N})$ and $E(sort(\frac{n_i}{N}))$ go to zero for most i.

Thus we only need to prove convergence of $sort(\frac{n_i}{N})$ to p' in L_1 in probability. This follows by an examination of the histogram order statistics $h_{n,j}$. Recall that these $h_{n,j}$ completely determine p_n. In addition, the $h_{n,j}$ satisfy a law of large numbers:

$$\frac{1}{m}h_{n,j} \to \frac{1}{m}E(h_{n,j}) \ \forall j \leq k,$$

for any finite k. (This can be proven, e.g., using McDiarmid's inequality.) Let us rewrite the above term more explicitly:

$$\frac{1}{m}E(h_{n,j}) = \frac{1}{m}\sum_{i=1}^{m} E(1(n_i = j)) = \frac{1}{m}\sum_{i=1}^{m} \binom{N}{j} p_i^j (1-p_i)^{N-j}.$$

Now we can rewrite this sum as an integral:

$$\frac{1}{m}\sum_{i=1}^{m} \binom{N}{j} p_i^j (1-p_i)^{N-j} = \int_0^1 dt \binom{N}{j} p_n(t)^j (1-p_n(t))^{N-j}, \qquad (1.20)$$

where

$$p_n(t) \equiv m p_{\max(i|i<t)}$$

is the discretized version of p. As the discretization becomes finer, the discretized version of p becomes close to p:

$$p_n \to p \ [\mu],$$

where $[\mu]$ denotes convergence in Lebesgue measure on the interval $[0,1]$ (this can be seen using approximation in measure by continuous functions of the almost surely finite function p). Since $N/m \to c$ and the integrand in equation (1.20) is bounded uniformly in N, we have by the dominated convergence theorem that

$$\int_0^1 dt \binom{N}{j} p_n(t)^j (1-p_n(t))^{N-j} \to \int_0^1 dt \frac{c^j}{j!} p(t)^j e^{-cp(t)}. \tag{1.21}$$

From the convergence of $h_{n,j}$ it easily follows that $p_n \to p'$ in L_1 in probability, where p' is determined by $E(h_{n,j})$ in the obvious way (since $p_n \to p'$ except perhaps on a set of arbitrarily small p'-measure). Since $\lim \frac{h_{N,0}}{m} > \int_0^1 dt I(p=0)$, $||p-p'||_1$ is obviously bounded away from zero.

Regarding the final claim of the theorem, the convergence of the $\hat{H}$ to $E(\hat{H})$ follows by previous considerations. We only need prove that $h(p') < h(p)$; after some rearrangement, this is a consequence of Jensen's inequality.

□

B.5 Sum inequalities

For $f(p) = 1/p$, we have the following chain of implications:

$$\begin{aligned} \sup_p |\frac{f(p)}{p}| &= c \\ \Rightarrow |f(p)| &\leq cp \\ \Rightarrow \sum_{i=1}^{m} f(i) &\leq c \sum p(i) = c. \end{aligned}$$

For the f described in section 1.6.1,

$$f(p) = \begin{cases} m & p < 1/m, \\ 1/p & p \geq 1/m, \end{cases}$$

we have

$$\sup_p |f(p)g(p)| = c \Rightarrow \sum_i g(i) \leq 2c,$$

since

$$\sum_i g(i) = \sum_{i:im \geq 1} g(i) + \sum_{i:im<1} g(i);$$

the first term is bounded by c, by the above, and the second by $cm\frac{1}{m} = c$.

This last inequality gives a proof of Proposition 5.

Statement (Proposition 5).

$$\max_p \sigma^2 \sim (\log m)^2.$$

Proof. We have $\max_p \sigma^2(p) = O(\log(m)^2)$: plug in $g = p\log(p)^2$ and take the maximum of fg on the interval. To see that in fact $\max_p \sigma^2(p) \sim (\log(m)^2)$, simply maximize $\sigma^2(p)$ on the central lines (section 1.7). □

B.6 Asymptotic bias rate

Statement (Theorem 9). *If* $m > 1$, $N \min_i p_i \to \infty$, *then*

$$\lim \frac{N}{m-1} B(\hat{H}_{MLE}) = -\frac{1}{2}.$$

Proof. As stated above, the proof is an elaboration of the proof of theorem 10.3.1 of [Devore and Lorentz, 1993]. We use a second-order expansion of the entropy function:

$$H(t) = H(x) + (t-x)H'(x) + (t-x)^2(\frac{1}{2}H''(x) + h_x(t-x)),$$

where h is a remainder term. Plugging in, we have

$$-t\log t = -x\log x + (t-x)(-1-\log x) + (t-x)^2(\frac{1}{2}\frac{-1}{x} + h_x(t-x)).$$

After some algebra,

$$(t-x)^2 h_x(t-x) = t - x + t\log\frac{x}{t} + \frac{1}{2}\frac{1}{x}(t-x)^2.$$

After some more algebra (mostly recognizing the mean and variance formulae for the binomial distribution), we see that

$$\lim_N N(B_N(H)(x) - H(x)) = H''(x)\frac{x(1-x)}{2} + R_N(x),$$

where

$$R_N(x) \equiv \frac{1-x}{2} - N\sum_{j=0}^{N} B_{j,N}(x)\frac{j}{N}\log\frac{j}{Nx}.$$

The proof of theorem 10.3.1 in [Devore and Lorentz, 1993] proceeds by showing that $R_N(x) = o(1)$ for any fixed $x \in (0,1)$. We need to show that

$R_N(x) = o(1)$ *uniformly* for $x \in [x_N, 1 - x_N]$, where x_N is any sequence such that

$$Nx_N \to \infty.$$

This will prove the theorem, because the bias is the sum of

$$B_N(H)(x) - H(x)$$

at m points on this interval. The uniform estimate essentially follows from the delta method (somewhat like Miller and Madow's original proof, except in one dimension instead of m): use the fact that the sum in the definition of $R_N(x)$ converges to the expectation (with appropriate cutoffs) of the function $t \log \frac{t}{x}$ with respect to the Gaussian distribution with mean x and variance $\frac{1}{N}x(1-x)$. We spend the rest of the proof justifying the above statement.

The sum in the definition of $R_N(x)$ is exactly the expectation of the function $t \log \frac{t}{x}$ with respect to the Binomial(N, x) distribution (in a slight abuse of the usual notation, we mean a binomial random variable divided by N, that is, rescaled to have support on $[0, 1]$). The result follows if a second-order expansion for $t \log \frac{t}{x}$ at x converges at an $o(1/N)$ rate in $Bin_{N,x}$-expectation, i.e., if

$$E_{Bin_{N,x}} N[t \log \frac{t}{x} - (t - x) - \frac{1}{2x}(t - x)^2] \equiv E_{Bin_{N,x}} g_{N,x}(t) = o(1),$$

for $x \in [x_N, 1 - x_N]$. Assume, wlog, that $x_N \to 0$; in addition, we will focus on the hardest case and assume $x_N = o(N^{-1/2})$. We break the above

expectation into four parts:

$$E_{Bin_{N,x}} g_{N,x}(t) = \int_0^{ax_N} gdBin_{N,x} + \int_{ax_N}^{x_N} gdBin_{N,x} + \int_{x_N}^{b_N} gdBin_{N,x} + \int_{b_N}^{1} gdBin_{N,x},$$

where $0 < a < 1$ is a constant and b_N is a sequence we will specify below. We use Taylor's theorem to bound the integrands near x_N (this controls the middle two integrals) and use exponential inequalities to bound the binomial measures far from x_N (this controls the first and the last integrals). The inequalities are due to Chernoff ([Devroye et al., 1996]): let B be $Bin_{N,x}$, and let a, b, and x_N be as above. Then

$$P(B < ax_N) < e^{aNx_N - Nx_N - Nx_N a \log a} \tag{1.22}$$

$$P(B > b_N) < e^{Nb_N - Nx_N - Nb_N \log \frac{b_N}{x_N}}. \tag{1.23}$$

Simple calculus shows that

$$\max_{t \in [0, ax_N]} |g_{N,x}(t)| = g_{N,x}(0) = \frac{Nx}{2}.$$

We have that the first integral is $o(1)$ iff

$$aNx_N - Nx_N - Nx_N a \log a + log(Nx_N) \to -\infty.$$

We rearrange:

$$aNx_N - Nx_N - Nx_N a \log a + \log(Nx_N) = Nx_N(a(1 - \log a) - 1) + \log(Nx_N).$$

Since

$$a(1 - \log a) < 1, \ \forall a \in (0, 1),$$

the bound follows. Note that this is the point where the condition of the theorem enters; if Nx_N remains bounded, the application of the Chernoff inequality becomes useless and the theorem fails.

This takes care of the first integral. Taylor's bound suffices for the second integral:

$$\max_{t\in[ax_N,x_N]} |g_{N,x}(t)| < |\max_{u\in[ax_N,x_N]} \frac{(t-x)^3}{-6u^2}|,$$

from which we deduce

$$|\int_{ax_N}^{x_N} gdBin_{N,x}| < \frac{N((1-a)x_N)^4}{-6(ax_N)^2} = o(1),$$

by the assumption on x_N.

The last two integrals follow by similar methods, once the sequence b_N is fixed. The third integral dies if b_N satisfies the following condition (derived, again, from Taylor's theorem):

$$\frac{N(b_N - x_N)^4}{-6x_N^2} = o(1),$$

or equivalently,

$$b_N - x_N = o(\frac{x_N^{1/2}}{N^{1/4}});$$

choose b_N as large as possible under this constraint, and use the second Chernoff inequality, to place an $o(1)$ bound on the last integral. □

B.7 Bayes concentration

Statement (Theorem 13). *If p is bounded away from zero, then H is normally concentrated with rate $m^{1/3}$; that is, for fixed a,*

$$p(|H - E(H)| > a) = O(e^{-Cm^{1/3}a^2}),$$

for any constant $a > 0$ and some constant C.

Proof. We provide only a sketch. The idea is that H almost satisifies the bounded difference condition, in the following sense: there do exist points $x \in [0,1]^m$ such that

$$\sum_{i=1}^{m} (\Delta H(x_i))^2 > m\epsilon_m^2,$$

say, where

$$\Delta H(x_i) \equiv \max_{x_i, x_i'} |H(x_1, \ldots, x_i, \ldots x_m) - H(x_1, \ldots, x_i', \ldots x_m)|,$$

but the set of such x — call the set A — is of decreasing probability. If we modify H so that $H' = H$ on the complement of A, and let $H' = E(H|p_i \in A^c)$ on A, that is,

$$H'(x) = \begin{cases} H(x) & x \in A^c, \\ \frac{1}{P(A^c)} \int_{A^c} P(x)H(x) & q \in A, \end{cases}$$

then we have that

$$P(|H' - E(H')| > a) < e^{-a^2(m\epsilon_m^2)^{-1}},$$

and

$$P(H' \neq H) = P(A).$$

We estimate $P(A)$ as follows:

$$\begin{aligned} P(A) &\leq \int_{[0,1]^m} 1(\max_i \Delta H(x_i) > \epsilon_m) d \prod_{i=1}^{m} p(x_i) \\ &\leq \int 1(\max_i (x_{i+2} - x_i) > \epsilon_m) dp^m(x_i) \\ &\sim \int dt e^{\log p - \epsilon_m pm}. \end{aligned} \tag{1.24}$$

The first inequality follows by replacing the L_2 norm in the bounded difference condition with an L_∞ norm; the second follows from some computation and the smoothness of $H(x)$ with respect to changes in single x_i. The last approximation is based on an approximation in measure by nice functions argument similar to the one in the proof of theorem 6, along with the well-known asymptotic equivalence (up to constant factors), as $N \to \infty$, between the empirical process associated with a density p and the inhomogeneous Poisson process of rate Np.

We estimate $|E(H) - E(H')|$ with the following hacksaw:

$$\begin{aligned} |E(H) - E(H')| &= |\int_A p(x)(H(x) - H'(x)) + \int_{A^c} p(x)(H(x) - H'(x))| \\ &= |\int_A p(x)(H(x) - H'(x)) + 0| \\ &\leq P(A) \log m. \end{aligned}$$

If $p > c > 0$, the integral in (1.24) is asymptotically less than $ce^{-m\epsilon_m c}$; the rate of the theorem is obtained by a crude optimization over ϵ_m. $\square$

The proof of the CLT (Theorem 14) follows upon combining previous results in this paper with a few powerful older results; again, to conserve space, we give only an outline. The asymptotic normality follows from McLeish's martingale CLT [Chow and Teicher, 1997] applied to the martingale $E(H|x_1, ...x_i)$; the computation of the asymptotic mean follows by methods almost identical to those used in the proof of Theorem 6 (sorting and linearity of expectation, effectively), and the asymptotic variance follows upon combining the formulae of [Darling, 1953] and [Shao and Hahn, 1995] with an approximation-in-measure argument similar, again, to that used to prove Theorem 6. See also [Wolpert and Wolf, 1995] and [Nemenman et al., 2002] for applications of Darling's formula to a similar problem.

B.8 Adaptive partitioning

Statement (Theorem 15). *If* $\log \Delta_N(\mathcal{A}_\mathcal{F}) = \mathrm{o}(\frac{\mathrm{N}}{(\log \mathrm{m})^2})$ *and* $\mathcal{F}$ *generates* $\sigma_{x,y}$ *a.s.,* $\hat{I}$ *is consistent in probability;* $\hat{I}$ *is consistent a.s. under the slightly stronger condition*

$$\sum \Delta_N(\mathcal{A}_\mathcal{F}) \mathrm{e}^{\frac{-\mathrm{N}}{(\log \mathrm{m})^2}} < \infty.$$

The key inequality, unfortunately, requires some notation; we follow the terminology in [Devroye et al., 1996], with a few obvious modifications. We take, as usual, $\{x_j\}$ as i.i.d random variables in some probability space $\Omega, \mathcal{G}, \mathrm{P}$ with empirical measure P_N. Let $\mathcal{F}$ be a collection of partitions of Ω,

with $\mathcal{P}$ denoting a given partition. $2^{\mathcal{P}}$ denotes, as usual, the "power set" of a partition, the set of all sets which can be built up by unions of sets in $\mathcal{P}$. We introduce the class of sets $\mathcal{A}_{\mathcal{F}}$, defined as the class of all sets obtained by taking unions of sets in a given partition, $\mathcal{P}$. In other words,

$$\mathcal{A}_{\mathcal{F}} \equiv \{\mathrm{A} : \mathrm{A} \in 2^{\mathcal{P}}, \mathcal{P} \in \mathcal{F}\}.$$

Finally, the Vapnik-Chervonenkis "shatter coefficient" of the class of sets $\mathcal{A}_{\mathcal{F}}$, $\Delta_N(\mathcal{A}_{\mathcal{F}})$, is defined as the number of sets which can be picked out of $\mathcal{A}_{\mathcal{F}}$ using N arbitrary points ω_j in Ω:

$$\Delta_N(\mathcal{A}_{\mathcal{F}}) \equiv \max_{\{\omega_\mathrm{j}\} \in \Omega^\mathrm{N}} |\{\omega_\mathrm{j}\} \cap \mathrm{A} : \ \mathrm{A} \in \mathcal{A}|.$$

The rate of growth in N of $\Delta_N(\mathcal{A}_{\mathcal{F}})$ provides a powerful index of the richness of the family of partitions $\mathcal{A}_{\mathcal{F}}$, as the following theorem (a kind of uniform LLN) shows; p here denotes any probability measure and p_N, as usual, the empirical measure.

Theorem (Lugosi and Nobel, '93). *Following the notation above, for any $\epsilon > 0$,*

$$P(\sup_{\mathcal{P} \in \mathcal{F}} \sum_{A \in \mathcal{P}} |p_N(A) - p(A)| > \epsilon) \leq 8\Delta_N(\mathcal{A}_{\mathcal{F}})\mathrm{e}^{-\mathrm{N}\epsilon^2/512}.$$

Thus this theorem is useful if $\Delta_N(\mathcal{A}_{\mathcal{F}})$ does not grow too quickly with N; as it turns out, $\Delta_N(\mathcal{A}_{\mathcal{F}})$ grows at most polynomially in N under various, easy-to-check conditions. Additionally, $\Delta_N(\mathcal{A}_{\mathcal{F}})$ can often be computed using straightforward combinatorial arguments, even when the number of

distinct partitions in $\mathcal{F}$ may be uncountable. See [Devroye et al., 1996] for a collection of instructive examples.

Proof. Theorem 15 is proven by a Borel-Cantelli argument, coupling the above VC inequality of Lugosi and Nobel with the following easy inequality, which states that the entropy functional H is "almost L_1 Lipshitz":

$$|H(p) - H(q)| \leq H_2(2||p - q||_1) + 2||p - q||_1 \log(m - 1),$$

where

$$H_2(x) \equiv -x \log(x) - (1 - x) \log(1 - x)$$

denotes the usual binary entropy function on $[0, 1]$. We leave the details to the reader. $\square$

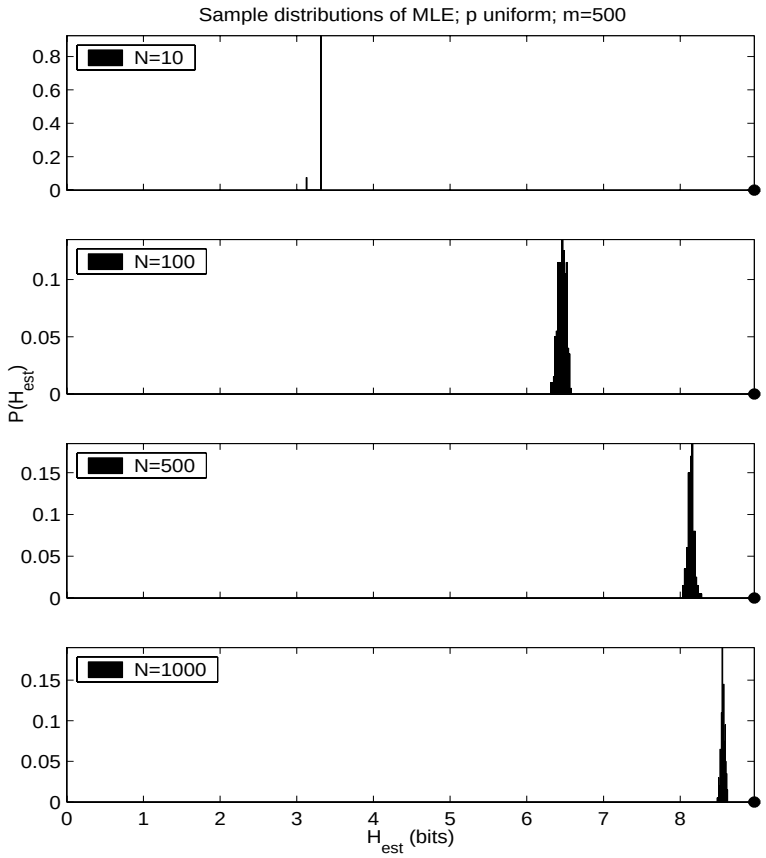

Figure 1.1: Evolution of sampling distributions of MLE: fixed m, increasing N. True value of H indicated by asterisk at bottom right corner of each panel. Note the small variance for all N, and the slow decrease of the bias as $N \to \infty$.

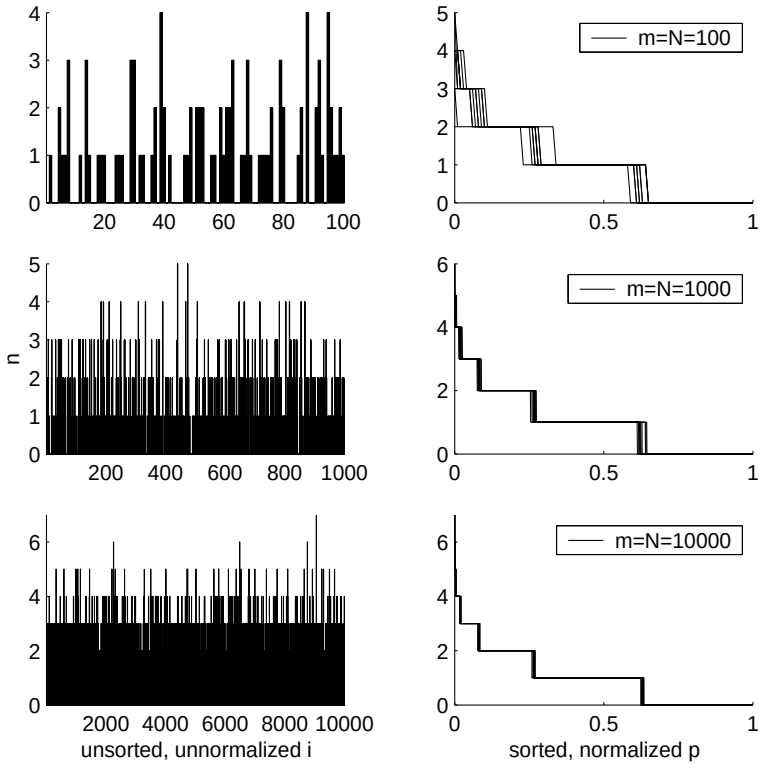

Figure 1.2: "Incorrect" convergence of sorted empirical measures. Each left panel shows an example unsorted m-bin histogram of N samples from the uniform density, with $N/m = 1$ and N increasing from top to bottom. Ten sorted sample histograms are overlaid in each right panel, demonstrating the convergence to a nonuniform limit. The analytically derived $p'_{c,\infty}$ is drawn in the final panel, but is obscured by the sample histograms.

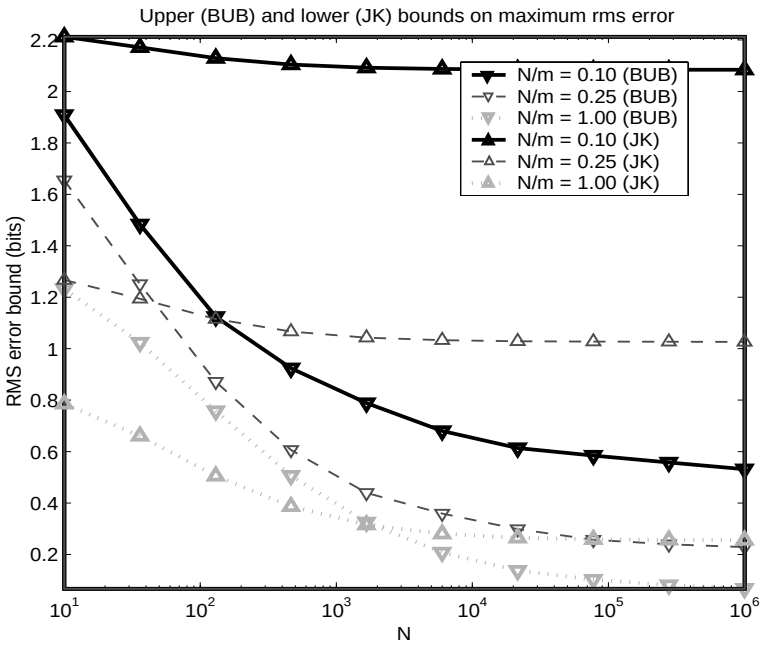

Figure 1.3: A comparison of lower bounds on worst-case error for $\hat{H}_{JK}$ (upward-facing triangles) to *upper* bounds on the same for $\hat{H}_{BUB}$ (down triangles), for several different values of N/m.

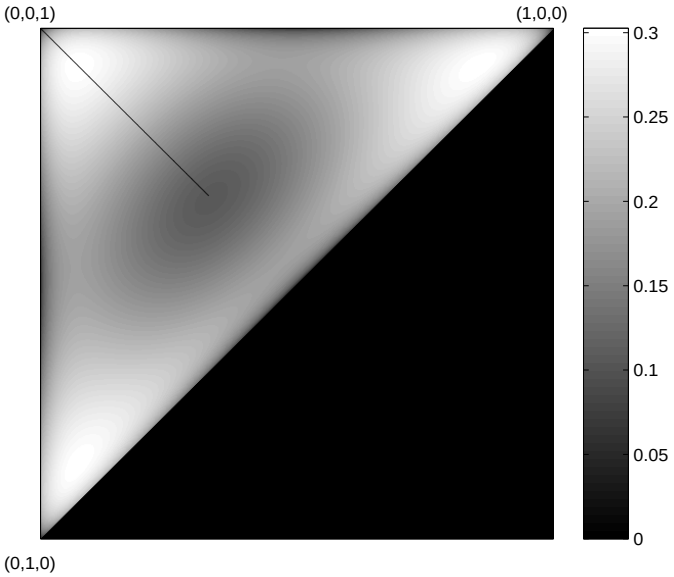

Figure 1.4: Exact RMS error surface (in bits) of the MLE on the 3-simplex, $N = 20$; note the six permutation symmetries. One of the "central lines" is drawn in black.

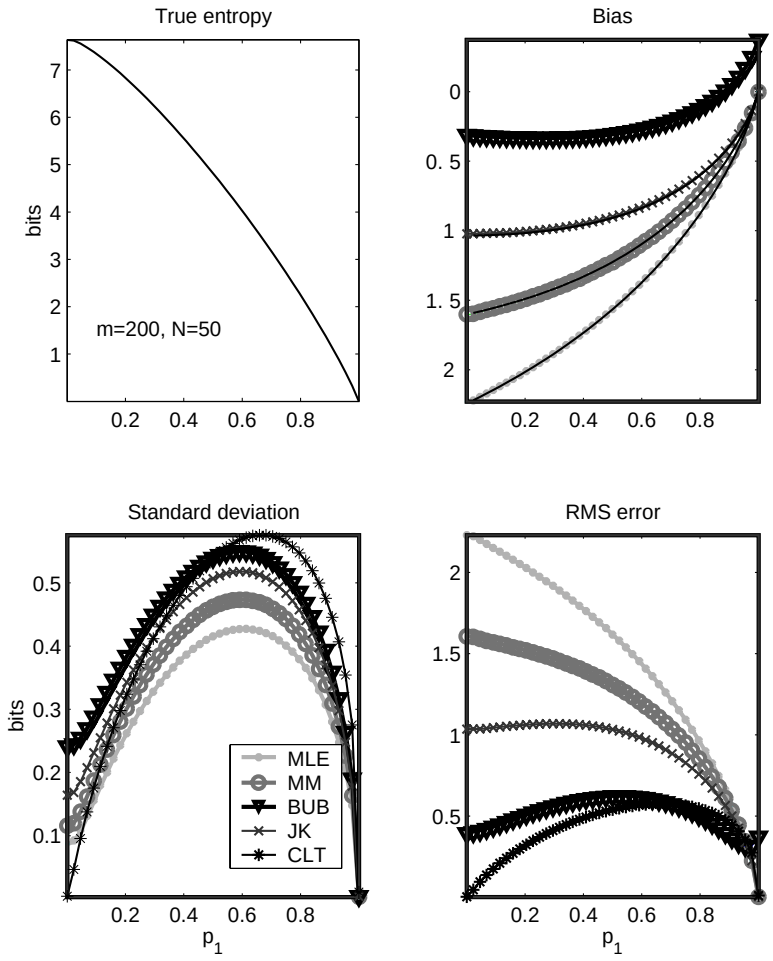

Figure 1.5: Example of error curves on the "central lines" for four different estimators ($N = 50, m = 200$; $\lambda_0 = 0$ here and below unless stated otherwise). The first panel shows the true entropy, as p_1 ranges from 1 (i.e., p is the unit mass on one point) to $\frac{1}{m}$ (where p is the flat measure on m points). Recall that on the central lines, $p_i = \frac{1-p_1}{m-1}$ $\forall i \neq 1$. The solid black lines overlying the colored (dashed or dotted) lines in the bias panel are the biases predicted by Theorem 6; these predictions depend on N and m only through their ratio, N/m. The black dash-asterisk denotes the variance predicted by the CLT, $\sigma(p)N^{-1/2}$.

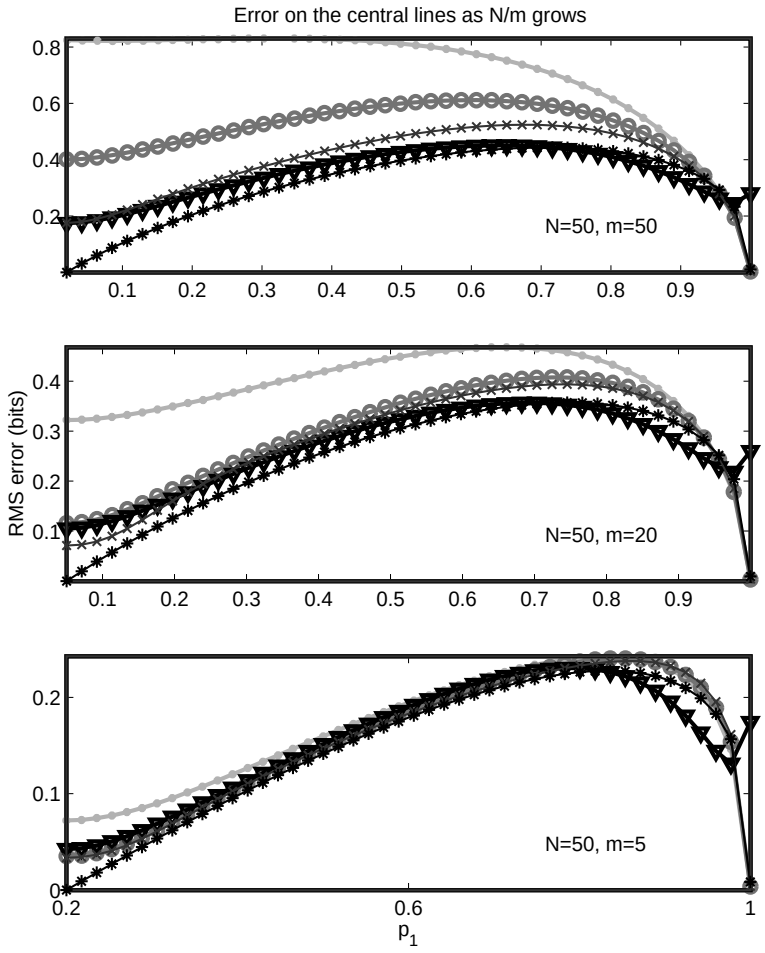

Figure 1.6: "Central line" error curves for three different values of N/m; notation as in Figure 1.5. Note that the worst-case error of the new estimator is less than that of the three most common $\hat{H}$ for all observed (N, m) pairs, and that the error curves for the four estimators converges to the CLT curve as $N/m \to \infty$.

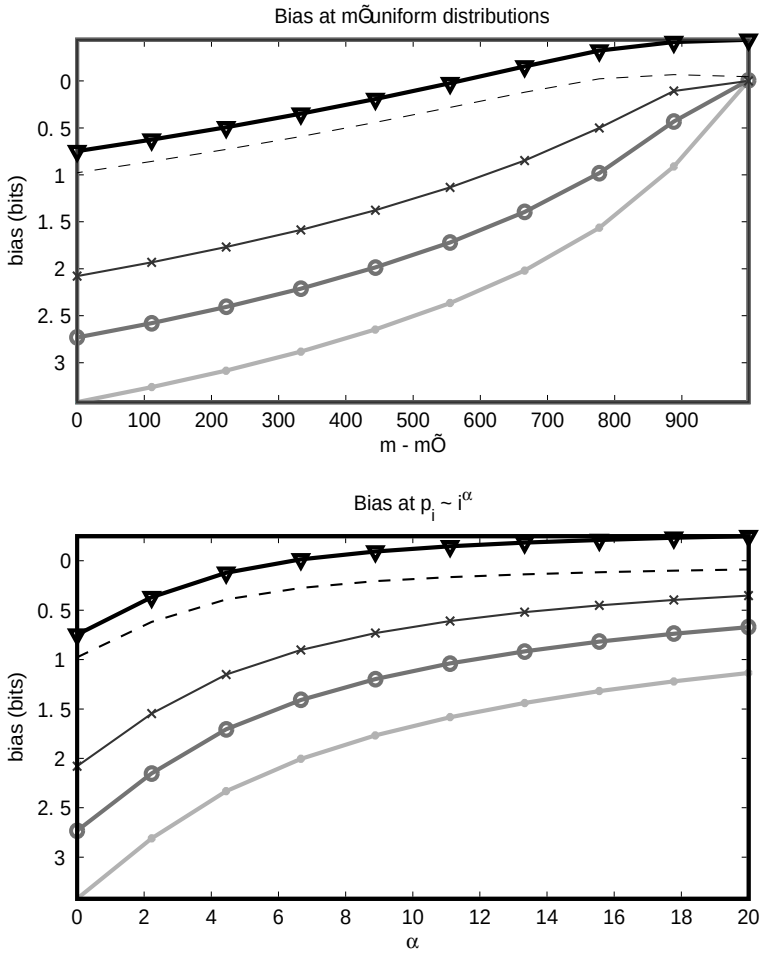

Figure 1.7: Exact bias on two additional families of distributions; notation as in Figure 1.5, plus dashed line corresponds to $\hat{H}_{BUB}$ with λ_0 set to reduce the bias at the zero-entropy point. Top panel: flat distributions on m' bins, $1 \leq m' \leq m$. Bottom panel: $p_i \simeq i^{\alpha}$. $N = 100$ and $m = 1000$ in each panel.

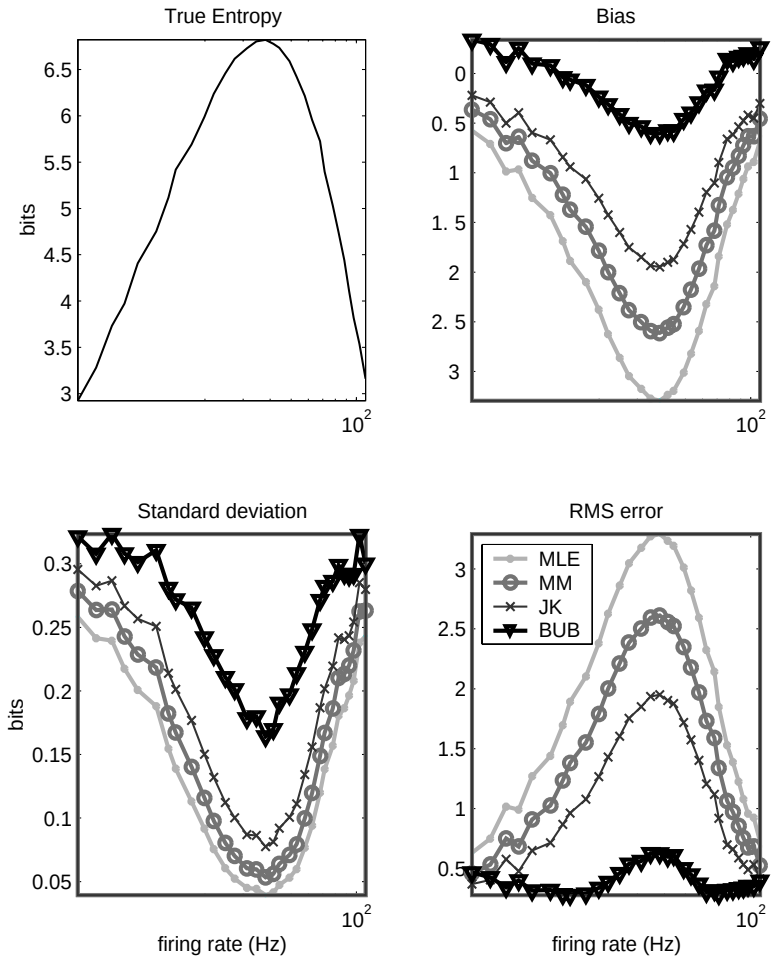

Figure 1.8: Error curves for simulated data (integrate and fire model, driven by white noise current), computed via Monte Carlo. $N = 100$ i.i.d. spike trains, time window $T = 200$ ms, binary discretization, bin width $dt = 20$ ms, thus, $m = 2^{10}$; DC of input current varied to explore different firing rates. Note the small variance and large negative bias of $\hat{H}$ over a large region of parameter space. The variance of $\hat{H}_{BUB}$ is slightly larger, but this difference is trivial compared to the observed differences in bias.

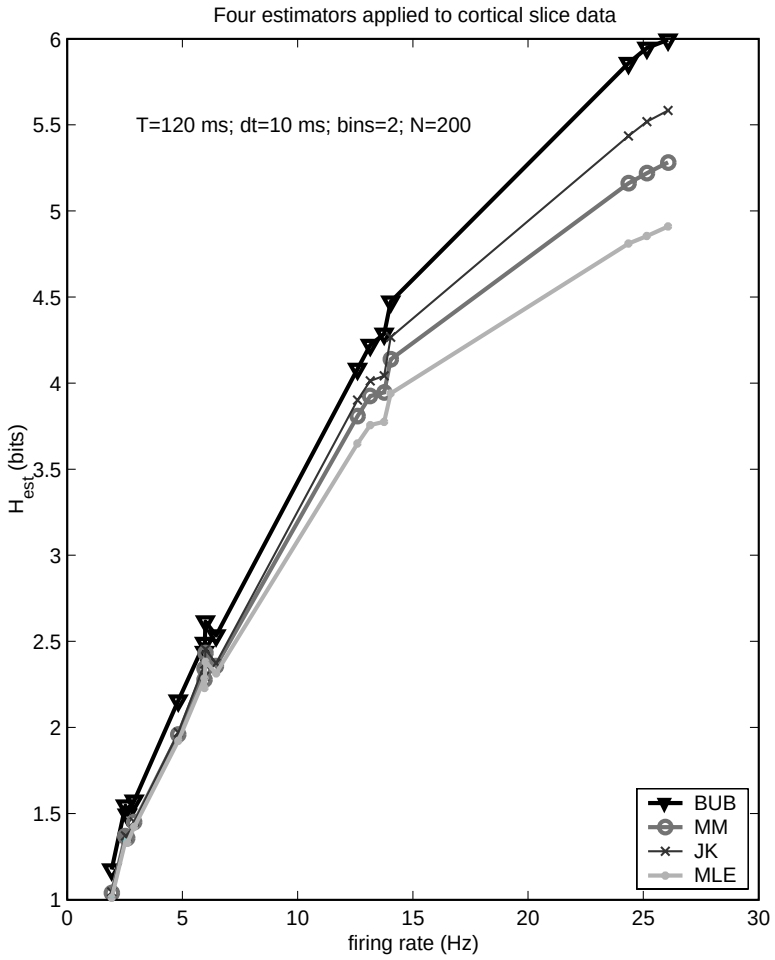

Figure 1.9: Estimated entropy of spike trains from a single cell recorded *in vitro*. Cell was driven with a white noise current. Each point corresponds to a single experiment, with $N = 200$ i.i.d. trials; the standard deviation of the input noise was varied from experiment to experiment. Spike trains were 120 ms long, discretized into 10 ms bins of 0 or 1 spike each; $m = 2^{12}$.

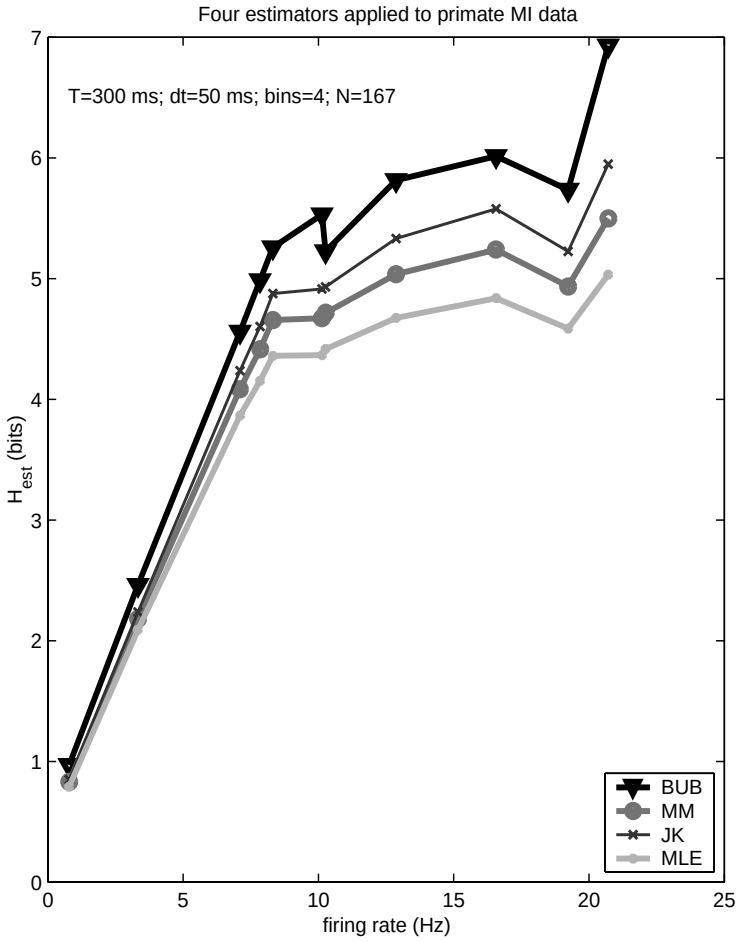

Figure 1.10: Estimated entropy of individual spike trains from 11 simultaneously recorded primate motor cortical neurons. Single units were recorded as monkey performed a manual random tracking task [Paninski et al., 1999, Paninski, 2003b]; N here refers to the number of trials (the stimulus was drawn i.i.d. on every trial). Each point on the x-axis represents a different cell; spike trains were 300 ms long, discretized into 50 ms bins of $0, 1, 2$, or > 2 spikes each; $m = 4^6$.

CHAPTER 2

Convergence properies of some spike-triggered analysis techniques

2.1 Introduction

Systems-level neuroscientists have a few favorite problems, the most prominent of which is the "what" part of the neural coding problem: what makes a given neuron in a particular part of the brain fire? In more technical language, we want to know about the conditional probability distributions $P(spike|X = x)$, the probability that our cell emits a spike, given that some observable signal X in the world takes value x. Because data is expensive, neuroscientists typically postulate a functional form for this collection of conditional distributions, and then fit experimental data to these functional models, in lieu of attempting to directly estimate $P(spike|X = x)$ for each possible x. Clearly, to interpret the results of this kind of statistical analy-

sis, we must have a good understanding of the bias and variance properties of the estimation procedure in question. This is especially true in the case of high-dimensional data (e.g., natural sensory signals or complex motor behavior), for which direct visualization is often impossible.

In this paper, we analyze the statistical properties of a phenomenological model whose popularity in the natural signal community seems to be on the rise [Theunissen et al., 2001, Brenner et al., 2001, Schwartz et al., 2002, Ringach et al., 2002]:

$$p(spike|\vec{x}) = f(<\vec{k}_1, \vec{x}>, <\vec{k}_2, \vec{x}>, \ldots, <\vec{k}_m, \vec{x}>). \tag{2.1}$$

Here f is some arbitrary $[0, 1]$-valued function, and $\{k_i\}$ are some linearly independent elements of (the dual space of) some vector space, X — the space of possible "input signals." Interpret f as a regular conditional probability. This model says, then, the neuron projects the signal $\vec{x}$ onto some m-dimensional subspace spanned by $\{\vec{k}_i\}_{1 \leq i \leq m}$ (call this subspace K), then looks up its probability of firing based only on this projection. This model is often called a "linear-nonlinear," or "LN," cascade model; it is a probabilistic analog of what are called "Wiener cascade" models [Hunter and Korenberg, 1986] in the system identification literature. (Note that this model is not the same as a Volterra series model [Marmarelis and Marmarelis, 1978]; these two classes of systems have very different approximation properties.)

The LN model has two important features which recommend it for com-

plex natural signal data. First, the spike trains of the cell are given by a conditionally (inhomogeneous) Poisson process given $\vec{x}$; that is, there are no dynamics in this model beyond those induced by $\vec{x}$ and K. This makes the LN cell a simple starting point for more detailed modeling. Second, equation (2.1) implies:

$$p(spike|\vec{x}) = p(spike|\vec{x} + \vec{y}) \; \forall \; y \perp K. \tag{2.2}$$

In other words, the conditional probability of firing is constant along (hyper)planes in the input space. This model thus separates the quite difficult nonparametric problem of learning $p(spike|\vec{x})$ into two much simpler pieces: learning K and learning f. For example, if f is known, the problem of learning K reduces to a fairly standard parametric estimation problem (for which, say, maximum likelihood methods are generally efficient); conversely, if K is known, learning f entails the nonparametric estimation of a density, about which, again, much is known (see e.g. [Devroye and Lugosi, 2001]). The semiparametric problem of estimating K without *a priori* knowledge of f seems to be much less well-understood; we focus primarily on this problem here.

Before we get to the heart of this paper, which is about estimating the parameters of this LN model, we should emphasize the model's phenomenological nature. As with all statistical models of nervous function, we are not attempting to describe the mechanism underlying a given cell's response properties, but rather to approximate this behavior with a model which is

both tractable from an estimation point of view, as emphasized above, and which also, we hope, provides some insight into the basic input-output properties of the cell. As others have argued, the LN model provides this type of tractable and understandable approximation in many different systems. We should also mention that the LN model can be justified, somewhat, using crude mechanistic ideas; for example, we could think of the nonlinearity as having to do with the spiking process, and the linear filtering as occurring in the dendrites, where we might hope that linearity might be a usable approximation. However, to date the statistical arguments for the model have tended to be much stronger than the mechanistic arguments (as, perhaps, the above "justification" demonstrates).

More generally, there are at least two goals of such a non-mechanistic analysis. First, most obviously, the more precise one's phenomenological description of a given system, the fewer choices one will have among mechanisms to explain the given behavior; thus statistical modeling can lead naturally into more detailed, mechanistic modeling (and vice versa). Second, statistical characterizations of a given system can be powerful even without deeper mechanistic understanding; for example, if one is interested not in how a message is encoded in the brain but rather in how well we can decode the message (this viewpoint is natural in the field of neural prosthetics [Serruya et al., 2002], for example), then good statistical theories are more valuable than mechanistic theories per se. In either case, it is essential to know the strengths and limitations of the statistical methods

used to estimate and model the properties of such systems, and this brings us back to the present work.

The main goal of this paper is to describe the convergence properties of three different types of K-estimator: 1) techniques based on the spike-triggered average [Theunissen et al., 2001]; 2) techniques based on the spike-triggered covariance [Brenner et al., 2001]; and 3) a new, more general technique based on a probabilistic distance measure between the spike-triggered and "no spike"- triggered distributions. We were motivated by two basic questions. First, when do these estimators work (in the sense of "consistency," that is, given enough data, do they provide us with an accurate estimate of K [Schervish, 1995])? Second, when the estimator is consistent, what is the statistical rate of convergence (that is, how much data do we need to be close to the correct K)? Our first main result is that the first two K-estimators often do not work, even given infinite data. More precisely, the conditions for consistency of these estimators turn out to be surprisingly stringent; for example, these conditions are typically not satisfied by natural signal stimulation paradigms. Our second main result provides an antidote of sorts: the novel estimator we introduce here converges under very general conditions to the correct K. Finally, we provide various results on the rate of convergence of these three classes of K-estimator. Together, these results serve to put the growing subfield of LN statistical modeling on a more solid theoretical foundation.

2.2 Notation; outline

The basic semiparametric model space we'll work in is defined as follows. An LN model is completely specified by knowledge of K and f, plus the stimulus distribution $p(\vec{x})$; thus, LN models take values in the space

$$(p, K, f) \in \mu(X) \times \mathcal{G}_m(X) \times L_{[0,1]}(\Re^m),$$

where $\mu(X)$ denotes the space of all probability measures on X, $\mathcal{G}_m(X)$ the space of all m-dimensional subspaces of X, and $L_{[0,1]}(\Re^m)$ the space of measurable functions on $\Re^m$ taking values in $[0, 1]$. In most cases, p is held fixed and/or presumed known, and we discuss only (K, f) instead of (p, K, f); this should be clear from context. Note that K, the main parameter of interest, is finite dimensional, while f, the "nuisance" parameter in statistical jargon, is infinite dimensional.

Let N denote the number of available samples, drawn i.i.d. from $p(\vec{x})$. We assume throughout this paper that $p(\vec{x})$ has zero mean and finite second moments; the first assumption obviously entails no loss of generality, and the second seems entirely reasonable on physical grounds. Then our basic results will take the following form:

$$E\left(\text{Error}(\hat{K})\right) \approx \alpha N^{-\gamma} + \beta, \tag{2.3}$$

as N becomes large. The estimator $\hat{K}$ is a map taking N observations of stimulus and spike data (where spikes are binary random variables, conditionally independent given the stimulus) into an estimate of the true un-

derlying K:

$$\begin{aligned} \hat{K} : (X \times \{0,1\})^N &\rightarrow \mathcal{G}_m(X) \\ (\vec{x}_N, s_N) &\rightarrow \hat{K}(\vec{x}_N, s_N), \end{aligned}$$

where $(\vec{x}_N, s_N)$ denotes the N-sample data; the natural error metric, then, is the geodesic distance on $\mathcal{G}_m(X)$ (the "canonical angle") between the true subspace K and the estimated subspace $\hat{K}$,

$$\text{Error}(\hat{K}) \equiv \cos^{-1}\left(s(P_K^t P_{\hat{K}})\right),$$

where P_V denotes the projection operator corresponding to the subspace V and $s(A)$ denotes the smallest singular value of the operator A. For notational ease, we will mostly work in the $m = 1$ case; here the metric takes the explicit form

$$\text{Error}(\hat{K}) \equiv \cos^{-1} \frac{< \hat{K}, \vec{k}_1 >}{||\hat{K}||_2 ||\vec{k}_1||_2}.$$

The scalar terms γ, α, and β in (2.3) each depend on f, K, and $p(\vec{x})$; γ is a constant giving the order of magnitude of convergence (usually, but not always, equal to 1/2), α gives the precise convergence rate, and β gives the asymptotic error. We will be mostly concerned with giving exact values for α, and simply indicating when β is zero or positive (i.e., when $\hat{K}$ is consistent in probability or not, respectively).

Most of the remainder of the paper will be devoted to deriving representation (2.3), including the constants α, β, and γ, for the three classes of

K-estimator mentioned in the introduction. We carry out this program for the spike triggered average and the spike-triggered covariance technique in sections 2.3 and 2.4, respectively. Section 2.5 contains perhaps the central results of this paper; here we give details on the analysis and computation of a new, universally consistent estimator. In section 2.6, we present some simulation results comparing the performance of the three estimators, and applications of the new estimator to real physiological data. We provide a few lower bounds on the convergence rate of any possible LN estimator in section 2.7; these bounds provide rigorous measures of the difficulty of the K-estimation problem. Finally (section 2.8), we close with a brief discussion of a few important areas for future research. Proofs appear in an appendix.

2.3 Spike-triggered averaging

The first estimator, the spike-triggered average, is classical and very intuitive. We define

$$\hat{K}_{STA} \equiv \frac{1}{N_s} \sum_{i=1}^{N_s} \vec{x}_i, \tag{2.4}$$

where $\vec{x}_i$ is the i-th stimulus for which a spike occurred and N_s denotes the total number of spikes observed (N and N_s are of course roughly proportional, with constant $p(spike) = \int_X p(\vec{x}) f(K\vec{x})$). As is well-known, $\hat{K}_{STA}$ is simply the sample mean of the spike-conditional stimulus distribution $p(\vec{x}|spike)$; since the spike signal is binary-valued, this is the

same as the cross-correlation between the spike and the stimulus signal. We will also consider the following linear regression-like modification [Theunissen et al., 2001]:

$$\hat{K}_{RSTA} \equiv A\hat{K}_{STA},$$

where A is an operator chosen to "rotate out" correlations in the stimulus distribution $p(\vec{x})$ (A is typically the (pseudo-) inverse of the stimulus correlation matrix, which we will denote as $\sigma^2(p(\vec{x}))$. In this section and the next, we assume that $\sigma^2(p(\vec{x}))$ is known; this assumption seems fair because either: 1) $p(\vec{x})$ is chosen by the experimenter, or, 2), in the natural signal paradigm, a sufficient number of samples from the natural distribution are available that $\sigma^2(p(\vec{x}))$ can be estimated to arbitrary accuracy; i.e., the experimenter has access to many more examples than N, the number of samples seen by the neuron. At any rate, even if $\sigma^2(p(\vec{x}))$ is unknown, the basic analysis presented here still works, although slightly worse constants are obtained).

We begin with necessary and sufficient conditions for consistency. As usual, we say $p(\vec{x})$ is radially symmetric if $p(B) = p(UB)$ for all measurable sets B and all unitary transformations (rotations) U; examples include the standard multivariate Gaussian density, or the uniform density on the sphere. (Note that if $p(\vec{x})$ has this radial symmetry property, then $\hat{K}_{STA} = \hat{K}_{RSTA}$.) Finally, since $\hat{K}_{RSTA}$ clearly returns a single vector, that is, a one-dimensional subspace of X, assume for the moment that $K = \vec{k}_1$

(i.e., K is a one-dimensional subspace). Then we have the following:

Theorem 21 (Consistency: $\beta(\hat{K}_{STA})$). *If $p(\vec{x})$ (resp. $p(A^{1/2}\vec{x})$) is radially symmetric and $E(<\vec{x},\vec{k}_1> | spike) \neq 0$, then $\beta(\hat{K}_{STA}) = 0$ (resp. $\beta(\hat{K}_{RSTA}) = 0$); that is, the spie-triggered average estimator is consistent.*

Conversely, if $p(\vec{x})$ is radially symmetric and $E(<\vec{x},\vec{k}_1> | spike) = 0$, then $\beta > 0$, and if $p(\vec{x})$ is not radially symmetric, then there exists an f for which $\beta > 0$.

In other words, spike-triggered averaging techniques always work (given enough data) if the input distribution p is radially symmetric, and if the neuron's tuning f is sufficiently asymmetric, in the sense that $|E(<\vec{x},\vec{k}_1> | spike)| > 0$; conversely, it is not hard to find examples for which these conditions are not met and the spike-triggered average fails to recover $\vec{k}_1$. The above sufficiency conditions are fairly well-known; for example, most of the sufficiency statement appeared (albeit in somewhat less precise form) in [Chichilnisky, 2001] (see also [Ringach et al., 1997] and references therein for related results; [Bussgang, 1952] seems to be the earliest). The condition on $E(<\vec{x},\vec{k}_1> | spike)$ is discussed in more depth below. Note the lack of restrictions on f; this function is not required to be smooth, or even continuous.

On the other hand, the converse is novel, to our knowledge, and is perhaps surprisingly stringent: without a highly restrictive symmetry condition on $p(\vec{x})$, spike-triggered averaging methods often remain biased,

even given infinite data; thus, these estimators will typically converge, but not necessarily to the correct $\vec{k}$. As is well known, distributions of natural signals tend to lack this symmetry property [Simoncelli, 1999, Ruderman and Bialek, 1994]; thus, spike-triggered average analyses of natural signal data must be interpreted with caution. The first part of the necessity statement will be obvious from the following discussion of $\alpha(\hat{K}_{RSTA})$ (and in fact appears implicitly in [Chichilnisky, 2001]). The second part, while perhaps unsurprising given the analysis of [Chichilnisky, 2001], is a little harder, and seems to require characteristic function (Fourier transform) techniques. The proof proceeds by showing that a distribution is symmetric iff it has the property that the conditional mean of $\vec{x}$ is zero on all planar "slices" (i.e., $E(< \vec{u}, \vec{x} > | < \vec{v}, \vec{x} > \in B) = 0$ for all $\vec{u} \perp \vec{v} \in X'$ and real measurable sets B).

Next we have the rate of convergence, to give a rough idea of how many samples is "enough":

Theorem 22 (Convergence rate: $\alpha(\hat{K}_{STA})$). *Assume $p(\vec{x})$ is independent normal, with standard deviation $\sigma(p)$. If $\beta(\hat{K}_{STA}) = 0$, then $N_s^{1/2}(\hat{K}_{STA} - K)$ is asymptotically independent normal with mean zero (considered as a distribution on the tangent plane of $\mathcal{G}_m(X)$ at the true underlying value K), and scale*

$$\frac{\sigma(p)}{E(< \vec{x}, \vec{k}_1 > | spike)}.$$

Thus,

$$\alpha(\hat{K}_{STA}) = \frac{\sigma(p)}{|E(<\vec{x}, \vec{k}_1 > |spike)|}\sqrt{\dim X - 1}.$$

Thus the performance of the spike-triggered average scales directly with the dimension of the ambient space and inversely with $|E(<\vec{x}, \vec{k}_1 > |spike)|$, a measure of the asymmetry of the spike-triggered distribution along k_1. The standard example of a neuron for which $|E(<\vec{x}, \vec{k}_1 > |spike)|$ is small is a complex cell in V1, whose responses are roughly symmetric with respect to sign inversion. The theorem serves to quantify the well-known result that spike-triggered averaging works poorly, if at all, for neurons with this kind of response symmetry.

The proof follows by applying the multivariate central limit theorem to the sample mean of N_s random vectors drawn i.i.d. from the spike-conditional stimulus distribution, $p(\vec{x}|spike)$. The proof also supplies the asymptotic distribution of $Error(\hat{K}_{STA})$ (a noncentral F), which might be useful for hypothesis testing. The details are easy once the mean of this distribution is identified (as in [Chichilnisky, 2001], under the above sufficiency conditions).

Note that we stated the result under stronger-than-necessary conditions (i.e., $p(\vec{x})$ is Gaussian instead of just symmetric), in order to simplify the statement. (In this case, the form of α becomes quite simple under these stronger assumptions; α depends on the nonlinearity f only through $E(<\vec{x}, \vec{k}_1 > |spike)$. The general case is proven by identical methods but results

in a slightly more complicated, f-dependent, term in place of $\sigma(p)$.) This pattern of stating non-optimal results in the text, then giving the stronger, more general results in the appendix, will reappear without comment below.

One final note: in stating the above two results, we have assumed that K is one-dimensional. Nevertheless, the two theorems extend easily to the more general case, after $Error(\hat{K}_{STA})$ is redefined to measure angles between $m-$ and $1-$dimensional subspaces. (Of course, now $E(\hat{K}_{STA})$ depends strongly on the input distribution $p(\vec{x})$, even for radially symmetric $p(\vec{x})$; see, e.g., [Schwartz et al., 2002] for an analysis of a special case of this effect.)

2.4 Covariance-based methods

The next estimator was introduced in an effort to extend spike-triggered analysis to the $m > 1$ case (see, e.g., [de Ruyter and Bialek, 1988, Brenner et al., 2001, Schwartz et al., 2002]). Where $\hat{K}_{STA}$ was based on the first moment of the spike-conditional stimulus distribution $p(\vec{x}|spike)$, $\hat{K}_{CORR}$ is based on the second moment. We define

$$\hat{K}_{CORR} \equiv (\sigma^2(p))^{-1}\mathrm{eig}(\widehat{\Delta\sigma^2}),$$

where $\mathrm{eig}(A)$ denotes the significantly non-zero eigenspace of the operator A, and $\widehat{\Delta\sigma^2}$ is some estimate (typically the usual sample covariance

estimate) of the "difference-covariance" matrix $\Delta\sigma^2$, defined by

$$\Delta\sigma^2 \equiv \sigma^2(p(\vec{x})) - \sigma^2(p(\vec{x}|spike)).$$

Again, we start with β:

Theorem 23 ($\beta(\hat{K}_{CORR})$). *If $p(\vec{x})$ is Gaussian and*

$$Var_{p(\vec{x}|spike)}(<\vec{k},\vec{x}>) \neq Var_{p(\vec{x})}(<\vec{k},\vec{x}>)\ \forall\vec{k} \in E_K,$$

for some orthogonal basis E_K of K, then $\beta(\hat{K}_{CORR}) = 0$. Conversely, if $p(\vec{x})$ is Gaussian and the variance condition is not satisfied for f, then $\beta > 0$, and if $p(\vec{x})$ is non-Gaussian, then there exists an f for which $\beta > 0$.

As before, the sufficiency is fairly well-known (see the thesis of Odelia Schwartz for a proof, or [Brenner et al., 2001] for a sketch), while the necessity appears to be novel and relies on characteristic function arguments. It is perhaps surprising that the conditions on p for the consistency of this estimator are even stricter than for the spike-triggered average. The essential fact here turns out to be that a distribution is normal iff, after a suitable change of basis, the conditional variance on all planar "slices" of the distribution is constant.

We have, with Odelia Schwartz, developed a striking inconsistency example which is worth mentioning here:

Example (Inconsistency of $\hat{K}_{CORR}$). *There is a nonempty open set of nonconstant f and radially symmetric $p(\vec{x})$ such that $\hat{K}_{CORR}$ is asymptot-*

ically orthogonal to K almost surely as $N \to \infty$. (In fact, the f and p in this set can be taken to be infinitely differentiable.)

The basic idea is that, for nonnormal p, the spike-triggered variance of $< \vec{v}, \vec{x} >$ depends on f even for $\vec{v} \perp \vec{k}$; thus, one can find f for which

$$|Var_{p(\vec{x}|spike)}(< \vec{k}, \vec{x} >) - Var_{p(\vec{x})}(< \vec{k}, \vec{x} >)|$$

is small but

$$|Var_{p(\vec{x}|spike)}(< \vec{v}, \vec{x} >) - Var_{p(\vec{x})}(< \vec{v}, \vec{x} >)|, \qquad \vec{v} \perp \vec{k},$$

is large. We leave the details to the reader.

We can derive a similar rate of convergence for these covariance-based methods. To reduce the notational load, we state the result for $m = 1$ only; in this case, we can define $\lambda_{\Delta\sigma^2}$ to be the (unique and nonzero by assumption) eigenvalue of $\Delta\sigma^2$.

Theorem 24 ($\alpha(\hat{K}_{CORR})$). *Assume $p(\vec{x})$ is independent normal. If $\beta(\hat{K}_{CORR}) = 0$, then $N_s^{1/2}(\hat{K}_{CORR} - K)$ is asymptotically independent normal with mean zero and*

$$\alpha = \frac{\sigma(p)\sqrt{\sigma^2(p) - \lambda_{\Delta\sigma^2}}}{|\lambda_{\Delta\sigma^2}|}\sqrt{\dim X - 1}.$$

(Again, while $\lambda_{\Delta\sigma^2}$ will not be exactly zero in practice, it can often be small enough that the asymptotic error remains prohibitively large for physiologically reasonable values of N_s.) The proof proceeds by applying the multivariate central limit theorem to the covariance matrix estimator, then

examining the first-order Taylor expansion of the eigenspace map at $\Delta\sigma^2$. It is also worth emphasizing that the asymptotics in the above theorem (and indeed, in all of the results in this paper) are in N only; the theorem is not valid if $\dim X$ grows as well. (See, e.g., [Everson and Roberts, 2000, Johnstone, 2000] and references therein for some useful asymptotic results on eigenspace analysis in the case that $\dim X$ is of order N.)

2.5 ϕ-divergence techniques

We have seen that the two most common K-estimators are not consistent in general; that is, the asymptotic error β is bounded away from zero for many (non-pathological) combinations of $p(\vec{x})$, f, and K. In particular, we have to place very strong conditions on p to guarantee that $\hat{K}_{RSTA}$ and $\hat{K}_{CORR}$ will converge to the correct K. We now introduce a new class of estimator which is consistent ($\beta = 0$) in great generality.

The basic idea is that $K\vec{x}$ is in a sense a sufficient statistic for $\vec{x}$: $\vec{x}$ — $K\vec{x}$ — *spike* forms a Markov chain. Let us give a few definitions. Given a continuous, strictly convex real function ϕ on $[0, \infty]$, with $\phi(1) = 0$, define the ϕ-divergence (following [Csiszar, 1967]) between two measures μ and ν as:

$$D_\phi(\mu; \nu) \equiv \int d\nu \phi(\frac{d\mu}{d\nu}) = \int d\mu \tilde{\phi}(\frac{d\nu}{d\mu}),$$

where $\tilde{\phi}(t) = t\phi(t^{-1})$ and the densities $d\mu$ and $d\nu$ are interpreted as likelihood ratios. The best-known ϕ-divergence is the Kullback-Leibler diver-

gence ($\phi(t) = t \log t$). The main property of ϕ-divergences we need is the so-called data-processing inequality [Cover and Thomas, 1991]: for any Markov morphisms S and T,

$$D_\phi(S(\mu); T(\nu)) \leq D_\phi(\mu; \nu),$$

with equality only if S and T are sufficient. The above inequality is named for the following special case: μ is $p(x, y)$, the joint distribution of some r.v.'s X and Y, and ν the product $p(x)p(y)$ of their marginals. Then, for any Markov chain X — Y — Z, $D_\phi(p(x, y); p(x)p(y)) \geq D_\phi(p(x, z); p(x)p(z))$, with equality iff X — Z — Y (i.e., iff $Y(Z)$ is sufficient for X).

Thus, if we identify the random variable *spike* with X in the above Markov chain, $\vec{x}$ with Y, and $< K, \vec{x} >$ with Z, it is clear from (2.1) that $< K, \vec{x} >$ is sufficient for $\vec{x}$ with respect to *spike*, and the data processing inequality states that

$$M_\phi(V) \equiv D_\phi\Big(p(< V, \vec{x} >, spike); p(< V, \vec{x} >)p(spike)\Big),$$

considered as a function of vector spaces V of dimension $\dim K$, reaches a maximum on K, and this maximum is unique under certain weak conditions. (When $\dim V > \dim K$, the maximum will no longer be unique, but it is easy to show that the maximizers still contain K.)

The basic idea is that $< V, \vec{x} >$ is equivalent to $< K, \vec{x} >$ plus some noise term that does not affect the spike process (more precisely, this noise term is conditionally independent of *spike* given $< K, \vec{x} >$); this noise term

is obviously 0 for $V = K$, and the larger this noise, the smaller $M_\phi(V)$. Another way to put it is that $M_\phi(V)$ measures how strongly $< V, \vec{x} >$ modulates the firing rate of the cell: for V near K, the conditional measures $p(spike| < V, \vec{x} >)$ are on average very different from the prior measure $p(spike)$, and $M_\phi(V)$ is designed to detect exactly these differences; conversely, for V orthogonal to K, the conditional measures $p(spike| < V, \vec{x} >)$ will appear relatively "unmodulated" (that is, $p(spike| < V, \vec{x} >)$ will tend to be much nearer the average $p(spike)$), and $M_\phi(V)$ will be comparatively small.

This all suggests that we could estimate K by maximizing $M_{\phi,N}(V)$, some estimator of the function $M_\phi(V)$. The rest of this section is devoted to describing the mathematical and computational properties of this type of estimator, for several different forms of $M_{\phi,N}(V)$. The precise choice of ϕ here seems not to matter much for the asymptotic analysis, as long as ϕ is smooth enough; for mathematical and computational convenience, we choose $\phi(t) = t^2 - 1$. For this ϕ, a little algebra shows that

$$M_\phi(V) = \frac{Var(p(spike = 1| < V, \vec{x} >))}{p(spike = 1)p(spike = 0)}$$

(where in a slight abuse of notation, $p(spike = 1)$ serves as a random variable — a function of $< V, \vec{x} >$ — in the numerator, and as a fixed probability in the denominator); this is a reasonably intuitive measure of firing rate modulation in the LN model. Originally, we chose this function because of the nice properties of ϕ and $t\tilde{\phi}(t)$ near zero, as previous work on

the estimation of mutual information [Paninski, 2003b] indicated that the smoothness of these functions plays a critical role in the estimability of D_ϕ; other advantages of this choice will become clear as we progress.

Before we move on to our main results, it is worth noting the recent work of [Sharpee et al., 2003], who independently presented an estimator based on maximizing the mutual information between $< V, \vec{x} >$ and *spike*; this corresponds, in our notation, to maximizing $M_\phi(V)$, with $\phi(t) = t \log t$. While the methods and analysis presented below are somewhat more detailed than and differ in several important respects from those described in [Sharpee et al., 2003], it is worthwhile consulting their work for another illustration of the improved performance of this kind of estimator. We hope to undertake a more thorough comparison of the statistical and computational efficiency of the two estimators in the future.

2.5.1 Asymptotics

We will start by defining $M_{\phi,N}(V)$ more precisely. The simpest idea would be to let $M_{\phi,N}(V)$ be a "plug-in" kernel or histogram estimator, that is,

$$M_{\phi,N}(V) \equiv D_\phi \left(\hat{p}_N(< V, \vec{x} >, spike); \hat{p}_N(< V, \vec{x} >) \hat{p}_N(spike) \right),$$

where $\hat{p}_N$, in turn, is an estimate of the underlying measure, either by kernel (that is, $\hat{p}_N$ is obtained by filtering the empirical measure

$$p_N \equiv \frac{1}{N} \sum_{i=1}^{N} \delta_i$$

according to some linear, shift-invariant kernel), or histogram (that is, X is partitioned into a countable number of bins, and $\hat{p}_N$ is simply the discrete measure induced by p_N). We denote such an estimator of K by

$$\hat{K}_\phi \equiv \text{argmax}_V M_{\phi,N}(V).$$

We assume that the chosen kernel or histogram partition is roughly isotropic, and that the data has been pre-whitened, so that the global scale of the data is roughly the same for every V; this helps to reduce the bias induced by the somewhat arbitrary scale imposed by the kernel width or average bin size. Fancier versions of these estimators adjust to the local scale as well (e.g., adaptive kernels or histograms), but for computational simplicity we will stick to nonadaptive estimators of the density for now. Obviously, for either type of estimator, we will have to let the kernel or bin width decrease with N; it is easy to come up with examples for which fixed bin width estimators fail (basically, because if the bin width is bounded from below, there exist f which are "averaged over" by the kernel or histogram). Thus much of the labor in the analysis of these estimators is in dealing with shrinking bin sizes.

Our first result is a general consistency result for the kernel estimator. A nearly identical result holds for the histogram estimator.

Theorem 25 ($\beta(\hat{K}_\phi)$). *If p has a nonzero density with respect to Lebesgue measure, f is not constant a.e., and the kernel width goes to zero more slowly than N^{r-1}, for some $r > 0$, then $\beta = 0$ for the kernel estimator.*

In other words, this new estimator $\hat{K}_\phi$ works for very general neurons f and stimulus distributions p; in particular, $\hat{K}_\phi$ is suitable for application to natural signal data. Clearly, the condition on f is minimal; we ask only that the neuron be tuned. The condition on p is quite weak (and can be relaxed further); we are simply ensuring that we are sampling from all of X, and in particular, the part of X on which the cell is tuned.

Things get more complicated when it comes to computing the rate of convergence. The rough picture is as follows: for each V, $M_{\phi,N}(V)$ converges to $M_\phi(V)$, with an error that depends on N, V, the kernel width a_N, and the parameters of the LN model (K, f, p). We have to choose a_N in such a way as to minimize the effect of these errors on $\hat{K}_\phi$. The error can be split up into a bias term and a variance term. It turns out that the variance term doesn't depend very strongly on a_N, so we ignore this for now. The bias term can be split up further into an approximation bias and a sample bias: the approximation bias measures the difference between $M_\phi(V)$ and its kernel- (or histogram-) smoothed version, defined in the obvious way, while the sample bias is the average difference between $M_{\phi,N}(V)$ and this smoothed version of $M_\phi(V)$. It is intuitively clear that these two types of bias behave differently as a function of a_N; if a_N goes to zero too slowly, the approximation bias will go to zero slowly but the sample bias will die quickly (roughly, because larger kernels or histogram bins average over more data), and vice versa. Thus, if we can compute the asymptotic approximation bias

and sample bias as a function of a_N, we have a well-defined optimization problem: choose a_N to minimize their sum, the total bias.

We carry out this program in the appendix; the final result, for $m = 1$, for example, is that the sample bias behaves roughly like $(Na_N)^{-1}$, implying that the naive estimator $\hat{K}_\phi$ converges somewhat slowly. The following theorem follows from some simple algebra to obtain the optimal kernel width for minimizing the bias in $M_{\phi,N}(V)$, then a second-order expansion of $E(M_{\phi,N}(V))$ around K to obtain the corresponding behavior for the bias of $\hat{K}_\phi$.

Theorem 26 (Bias of $\hat{K}_\phi$). *If the approximation error is of order a_N^r, then the optimal kernel width is of order $N^{\frac{-1}{r+1}}$, corresponding to an optimal bias in the kernel or histogram estimators which can be of order $N^{\frac{-r}{2(r+1)}}$.*

Again, a similar conclusion holds for the histogram version of $\hat{K}_\phi$. To understand what this result means for a given set of parameters (f, K, p), note that it is straightforward to show, using a Taylor expansion, that the approximation error behaves like a_N^2 if p is well-behaved and f is, say, uniformly twice differentiable; this corresponds to a convergence rate of $N^{-1/3}$ for $\hat{K}_\phi$. As another example, step functions have an approximation error that behaves like a_N; this leads to an even slower convergence rate, $N^{-1/4}$.

This slow bias behavior can be corrected using a standard statistical trick: we replace the naive "plug-in" estimators for $M_{\phi,N}$ with their jack-

knifed versions, where for any function of the data $T(x_N)$, we define the jackknifed version of T to be

$$T_{JK} = NT - \frac{N-1}{N}\sum_{i=1}^{N} T_{-i},$$

where T_{-i} is T computed using all but the i-th data sample. Simple computations prove that this procedure solves the bias problem (for simplicity, the next three results in this section are stated under some weak smoothness assumptions on f and p; see the appendix for details):

Proposition 27 (Jackknife bias). *If the kernel (or bin) width goes to zero more slowly than N^{r-1}, $r > 0$, then the sample bias of the jackknifed version of $M_{\phi,N}$ decays exponentially.*

This is almost enough to establish an $N^{-1/2}$ convergence rate for the estimator $\hat{K}_\phi$ given by maximizing the jackknifed kernel or histogram version of $M_{\phi,N}$, under suitable conditions on the smoothness of f. The last step is to show that $M_{\phi,N}(V)$ is asymptotically linear in N and smooth enough in V, that is,

$$M_{\phi,N}(V) = M_\phi(V) + p_N m_V + o_p(N^{-1/2}), \tag{2.5}$$

where $o_p(N^{-1/2})$ is a random variable which is negligible on an $N^{1/2}$ scale, and

$$p_N m_V \equiv \frac{1}{N}\sum_{i=1}^{N} m_V(\vec{x}_i, spike_i)$$

denotes the "empirical process" associated with some function $m_V(\vec{x}, spike)$, uniformly differentiable in V and with

$p(\vec{x}, spike)$-mean zero. We leave the details behind representation (2.5) for the appendix; basically, m_V is computed as a derivative of $M_{\phi,N}(V)$. Now, the theory of empirical processes [van der Vaart and Wellner, 1996] states that $p_N m_V$ converges in a suitable sense to a Gaussian stochastic process (this makes intuitive sense, given that for fixed V, $p_N m_V$ is just a sample mean of N i.i.d. random variables with finite variance), and this leads, finally, to the asymptotic representation for $\hat{K}_\phi$:

Theorem 28 (γ and α for $(\hat{K}_\phi)$). *If the approximation error is of order a_N^r, $r > 1$, then the jackknifed kernel or histogram versions of $\hat{K}_\phi$, with bandwidth N^s, $-1 < s < -1/r$, converge at an $N^{-1/2}$ rate.*

Moreover, $N^{1/2}(\hat{K}_\phi - K)$ is asymptotically normal, with mean zero and

$$\alpha(\hat{K}_\phi) = (\text{trace } H^{-1} J H^{-1})^{1/2},$$

where

$$H \equiv \left. \frac{\partial^2 M(V)}{\partial V^2} \right|_K$$

and

$$J \equiv E_{p(\vec{x},spike)} \left(\left. \frac{\partial m_V}{\partial V} \right|_K^2 \right).$$

The methods follow, e.g., example 3.2.12 of [van der Vaart and Wellner, 1996] — basically, a generalization of the classical theorem on the asymptotic distribution of the maximum likelihood estimator in regular parametric families.

Numerical evidence indicates that $\alpha(\hat{K}_\phi)$ is often smaller than $\alpha(\hat{K}_{STA})$ or $\alpha(\hat{K}_{CORR})$ (that is, the ϕ-divergence estimator often converges faster than

the spike-triggered average or covariance methods, even in the cases when the latter two methods are known to converge to the correct K), but we have so far been unable to obtain any general bounds on these quantities. Section 2.6 details a few of these numerical experiments, using both simulated and real data; see also [Sharpee et al., 2003] for some simulations of a similar estimator using natural image data.

2.5.2 Computation

We still haven't mentioned how to actually compute $\hat{K}_\phi$. Histogram methods for the evaluation of $M_{\phi,N}(V)$ suffer from several problems: it is difficult to non-adaptively place histogram partitions well for all V simultaneously, for example, and attempts to place the histogram adaptively greatly complicate hill-climbing algorithms for the maximization of $M_N(V)$. Kernel methods are more attractive, but require numerical integration of effectively unconstrained nonlinear functions over m-dimensional spaces. A more efficient approach is a "resubstitution" estimator: we replace numerical integration with a kind of Monte Carlo integration, using the observed samples as our integration points. Thus, sticking with the example of $\phi(t) = t^2 - 1$, instead of computing the integral

$$\int \frac{\hat{p}(V\vec{x}, spike)^2}{\hat{p}(V\vec{x})\hat{p}(spike)} = \int \hat{p}(spike|V\vec{x})\hat{p}(V\vec{x}|spike),$$

we compute the sum

$$\frac{1}{N_s}\sum_{i\in S}\hat{p}(spike=1|V\vec{x}_i)+\frac{1}{N-N_s}\sum_{i\in S^c}\hat{p}(spike=0|V\vec{x}_i), \tag{2.6}$$

where S is the set of stimuli which induced a spike, and ^{c} denotes the set complement. The conditional measures $\hat{p}(spike|V\vec{x})$ are estimated via kernel, as discussed above; again, the jackknife trick can be used to remove the sample bias, and the asymptotic theory developed in the last section goes through.

To compute $\hat{K}_\phi$, now, we have to maximize $M_{\phi,N}(V)$; unfortunately, this function is non-convex in general, and no direct solution seems to exist. General iterative algorithms such as simulated annealing or gradient ascent with repeated restarts may, of course, be applied to this problem, but their convergence is extremely slow. We have developed a specialized ascent algorithm for maximizing expression (2.6) that is much more efficient. This algorithm makes use of several tricks which might be useful more generally for maximizing empirical functionals on spaces of vector spaces; we describe these ideas in turn below. We plan to make the algorithm publicly available at http://www.cns.nyu.edu/∼liam, in order to facilitate quantitative evaluation on as large a variety of neural and synthetic data as possible.

The basic algorithm alternates between a local step and a global step until a convergence criterion is satisfied. The local step is straightforward: given the current V_0, we compute the gradient of (2.6), using a smooth (Gaussian, say) kernel; call the gradient $\vec{e}_0 \perp V_0$. The global step consists of

finding the constrained maximum of (2.6), where V is allowed to vary only over the circle

$$(1+t^2)^{-1/2}(V_0 + t\vec{e}_0),\ t \in \Re. \tag{2.7}$$

The first, and most important, trick now is to compute $M_{\phi,N}(V)$ using not a smooth kernel, but rather a simple boxcar function. This allows us to compute our function for all t very quickly (and therefore to find its global maximum over all t very quickly (noniteratively) as well. The idea is simple: for a boxcar kernel, $M_{\phi,N}(t)$ changes value a finite number of times, namely at the points t_i at which kernels centered on different points intersect. Since precomputing these "crossing times" t_i is simple trigonometry, we only need to sort the times and keep track of the value of each change (this turns out to be very simple as well, since (2.6) is a sum over the same index i) to compute the full function. This mix of global and local maximizations greatly increases the efficiency of the algorithm. This also obviates the need for conjugate gradient ascent techniques [Press et al., 1992], as the boxcar kernels make $M_{\phi,N}(V)$ highly unsmooth (i.e., we don't become trapped in any long, smooth valleys).

Our other tricks do not have quite the same impact, but are helpful nevertheless. The next two ideas are about choosing the search direction $\vec{e}_0$ intelligently when local maxima are encountered (i.e., when the circle search described above returns V_0 as the global maximum over t). First, if we have kept a list of circles we have already searched over, we can use

a self-avoiding procedure to choose our next search direction: basically, we choose the next search to be in the direction $\vec{e}_0$ such that

$$\vec{e}_0 = \mathrm{argmax}_{\vec{e} \perp V_0} \max_z D\left(V_z \oplus \vec{e}_z, V_0 \oplus \vec{e}_0 \right),$$

where $D()$ denotes geodesic distance and z indexes all past searches. This prevents us from searching over ground we have already covered.

The second trick along these lines is a little more interesting, but requires that at least $\dim X$ searches have already been made (roughly, we will need the set of old search circles to span X before this method becomes useful). The idea is that, with each search, we gain some information on the global structure of $M_{\phi,N}(V)$, beyond the simpler local structure we use to choose gradients, do circle maximizations, and so on. If we can use this global information to guide our choice of the next search direction, we should gain in efficiency. The simplest way to do this is a variant of the principal component analysis-style trick used by the spike-triggered covariance estimator. We form two "covariance" matrices, U and V, as follows: U is the covariance of a set of points $\vec{y}_i$ sampled randomly from the set of all previous search circles (this is a set of $\dim X$-dimensional points, all of length unity), and V is the covariance of the same set of points, with norm scaled now by the value of $M_{\phi,N}(V)$ at each point, i.e., $M_{\phi,N}(\vec{y}_i)\vec{y}_i$. By the unitary symmetry of $M_{\phi,N}(V)$, we can hope that the "variance" of the data $M_{\phi,N}(\vec{y}_i)\vec{y}_i$ should be largest near K, even if we haven't searched (i.e., collected points $\vec{y}_i$) near K yet. Now our best guess at a good search direc-

tion $\vec{e}_0$ solves the usual eigenvector problem associated with the Rayleigh quotient corresponding to U and V.

Finally, we can use a few not-so-specialized tricks to help speed up convergence. Most of these are some version of the coarse-to-fine idea. Since the speed of the algorithm scales inversely with N, but the accuracy scales proportionally with N, we can run the algorithm for a few iterations on a subsampled data set (artificially reducing N) to get a rough estimate, then gradually scale up N to refine our original coarse estimate. Similar tricks can be played with the kernel width a and $\dim X$, assuming f and K, respectively, vary slowly enough that coarsening makes sense.

2.6 Application to simulated and real data

In this section we give examples of data sets, both simulated and real, for which the novel estimator introduced in section 2.5 reveals structure that is either undetected or contaminated by the usual estimators $\hat{K}_{STA}$ and $\hat{K}_{CORR}$. See, e.g., [**?**, Sharpee et al., 2003] for further numerical comparisons.

2.6.1 Numerical comparisons

Figures 2.1-2.3 present simple comparisons of the performance of $\hat{K}_\phi$ to that of the standard estimators $\hat{K}_{STA}$ and $\hat{K}_{CORR}$ on simulated data. Simulations here have the advantage, as usual, that we know the "right answer";

this allows us to rigorously quantify the distribution of error of these estimators in simple, easy-to-understand situations, and to illustrate, in a less technical way, some of the ideas presented in more mathematical language in the preceding sections. Each point in each of these first three figures corresponds to the error of the two estimators ($\hat{K}_\phi$ versus $\hat{K}_{STA}$ in 2.1 and 2.2, and versus $\hat{K}_{CORR}$ in Fig. 2.3), given N samples drawn i.i.d. from a fixed distribution $p(\vec{x})$ and presented to an LN model whose parameters were chosen randomly on each set of N trials. In each case, the LN model is one-dimensional ($m = 1$), for simplicity.

In Fig. 2.1, we chose the input distribution $p(\vec{x})$ and the parameters of the LN model to be entirely favorable to the performance of $\hat{K}_{STA}$: $p(\vec{x})$ was chosen to be standard Gaussian to satisfy the conditions of Theorem 21 (implying that the spike-triggered average does not suffer from an asymptotic bias), while the nonlinearity f was chosen to be a simple Heaviside step function (taking values zero and one), where the step position was chosen randomly according to a standard normal as well (by Theorem 22, this form of f implies that $\alpha(\hat{K}_{STA})$ is always finite, and indeed fairly small with high probability; the value of the linear filter $\vec{k}$ is irrelevant, by the symmetry of p). Nevertheless, somewhat surprisingly, $\hat{K}_\phi$ significantly outperforms $\hat{K}_{STA}$ on average ($p < .05$, rank test).

We chose the LN model parameters randomly in Fig. 2.1, partly in an effort to emulate physical reality, where we have no control over the parameters, and partly to avoid picking an LN model that happened to

confound either estimator to an abnormal degree. However, it is worth showing an example of the estimators' performance on a single, fixed model and input distribution, both because single models are perhaps easier to think about than a family of random models and in order to give a sense of the variability involved in the above numerical experiment. Thus, in Fig. 2.2, we present an identical simulation, except with the position of the step in the nonlinearity f (the only random parameter) fixed at 0. The results are essentially identical, if anything favoring the new estimator even more.

In Fig. 2.3, we use a nonlinearity which is more suited to $\hat{K}_{CORR}$: f is quadratic, of the form

$$f(t) = a(t-b)^2,$$

with a, b chosen randomly ($a > 0$; we have in mind an energy-type model for visual cortex cells; see, e.g., [Simoncelli and Heeger, 1998] and references therein). For Gaussian input data, $\hat{K}_{CORR}$ is competitive with $\hat{K}_{\phi}$ (data not shown), as expected given theorem 23. To provide a physiologically plausible example for which this is not the case, we took the input distribution $p(\vec{x})$ to be uniform on a hypercube; this corresponds, for example, to a temporal signal whose value is chosen independently and identically distributed at each time step. The physical examples we have in mind here are the full-field white visual flicker stimulus employed, for example, in [Berry and Meister, 1998, Chander and Chichilnisky, 2001], or the random "checkerboard" spatial stimulus used in cortical and thalamic studies

(e.g., [Reid and Shapley, 2002] and references therein). Finally, we chose the linear projection $\vec{k}$ randomly on the sphere for each new set of data; the results did not depend strongly on the identity of the chosen filter (for instance, the ratio $Error(\hat{K}_{CORR})/Error(\hat{K}_{\phi})$ was uncorrelated with the smoothness of $\vec{k}$; data not shown). Figure 2.3 shows that the new estimator outperforms the covariance-based estimator by a wide margin, essentially because of the asymptotic bias effects caused by the non-Gaussianity of the data, as discussed in Theorem 23.

2.6.2 Retinal ganglion cell data

We turn now to a simple application to real data. We use a data set described in detail in [Chander and Chichilnisky, 2001] (see also, e.g., [Schwartz et al., 2002, Pillow and Simoncelli, 2003]): in brief, salamander retinal ganglion cells were recorded extracellularly in vitro while a white binary full-field flicker visual signal (update rate = 33 Hz) was presented. For this data set, it turns out that $\hat{K}_{\phi}$ and $\hat{K}_{STA}$ are consistently highly correlated (assuming $\hat{K}_{\phi}$ is allowed to search for only one vector; data not shown), and therefore the performance of these estimators is quite similar.

More recently, interest has focused on less informative subspaces, specifically on "suppressive axes" [Schwartz et al., 2002], which are, for example, revealed by the secondary eigenstructure of $\hat{K}_{CORR}$, but not detected by standard spike-triggered average analysis (these secondary axes are, by definition, orthogonal to $\hat{K}_{STA}$). However, as discussed in section 2.4, $\hat{K}_{CORR}$

can be heavily biased if the input distribution $p(\vec{x})$ does not satisfy certain conditions (c.f. Theorem 23). The current distribution $p(\vec{x})$ is strongly non-Gaussian, and therefore certainly does not satisfy these conditions ($p(\vec{x})$ is not even elliptically symmetric, the weaker condition of Theorem 21). As we show in Fig. 2.4 (and in Fig. 2.3, as well), this can cause some problems. The left two panels here show $p(< \hat{K}_{CORR}, \vec{x} >)$ and $p(spike \mid < \hat{K}_{CORR}, \vec{x} >)$ — the distribution of the raw data projected onto the secondary axis of $\hat{K}_{CORR}$, and an estimate of the firing rate given this projection, respectively — for a single ganglion cell.

We see in the bottom panel that $\hat{K}_{CORR}$ has picked out a somewhat bizarre-looking, nearly discrete projection of $p(\vec{x})$. What, exactly, do we mean by "bizarre," here? As discussed in [Diaconis and Freedman, 1984], random projections of high-dimensional data are nearly Gaussian with high probability (given the right definitions of "random projection," "high-dimensional," and"nearly Gaussian," of course, and under certain weak conditions which happen to be satisfied by the independent binary $p(\vec{x})$); therefore, when an ostensibly physical projection, which should average over many independent time bins, and which should therefore (speaking roughly) qualify for consideration under the theorem of [Diaconis and Freedman, 1984], looks this non-Gaussian, it can be taken as a sign that something is wrong. In fact, this phenomenon was predicted by Theorems 21 and 23 (see also [Chichilnisky, 2001]): $\hat{K}_{CORR}$ is being pulled in the directions of the "corners" of the hypercube from which $p(\vec{x})$

is sampling, and is thus strongly biased away from the physically relevant axes that are modulating the cell's activity.

What happens when we estimate these secondary axes using $\hat{K}_\phi$ (constrained to return an axis orthogonal to $\hat{K}_{STA}$), instead of $\hat{K}_{CORR}$? According to the discussion in section 2.5, the new estimator should not be susceptible to the artifacts biasing $\hat{K}_{CORR}$, and as far as we can tell without objective knowledge of the "true" underlying suppressive axes for this cell, this appears to be the case. In the right panels of Fig. 2.4, we see that $p(< \hat{K}_\phi, \vec{x} >)$ looks much more Gaussian than $p(< \hat{K}_{CORR}, \vec{x} >)$, for example. More directly, $\hat{K}_\phi$ captures significantly ($\approx 50\%$) more information about the cell's firing behavior than does $\hat{K}_{CORR}$, despite the fact that $\hat{K}_\phi$ was trained on only 10% of the data provided to $\hat{K}_{CORR}$. Thus, $\hat{K}_\phi$ appears to outperform $\hat{K}_{CORR}$ on real data as well, at least in the not uncommon case that the input distribution is non-Gaussian.

2.6.3 Motor cortical data

In the preceding, we provided some encouraging numerical comparisons between $\hat{K}_{STA}$, $\hat{K}_{CORR}$, and the new estimator $\hat{K}_\phi$. This last subsection presents some preliminary results which are of interest more for their physiological relevance than for methodological reasons.

We have begun to apply these new spike-triggered analysis techniques to data collected in the primary motor cortex (MI) of awake, behaving monkeys, in an effort to elucidate the neural encoding of

time-varying hand position signals in MI. This analysis has led to several interesting findings on the encoding properties of these neurons, with immediate applications to the design of neural prosthetic devices [Paninski et al., 2002, Shoham et al., 2003]. The monkeys are performing a random drawing task, designed roughly to mimic everyday (for humans, but perhaps not monkeys in the wild) manual movement (for methodological details, see [Paninski et al., 1999, Fellows et al., 2001, Paninski et al., 2003a, Serruya et al., 2002]); the "stimulus" space X in this context is the fairly high-dimensional space of time-varying hand position signals.

One novel and surprising result of this analysis is that the relevant K for MI cells appear to be one-dimensional. In other words, the conditional firing rate of these neurons, given a specific time-varying hand path, is well captured by the following model (Fig. 2.5): $p(spike|\vec{x}) = f(< \vec{k}_1, \vec{x} >)$, where $\vec{x}$ represents the two-dimensional hand position signal in a temporal neighborhood of the current time, $\vec{k}_1$ (in a slight abuse of notation) is a cell-specific affine functional, and $f(t)$ is a scalar nonlinearity which turns out to be relatively cell-independent. There is no reason to have assumed MI cells would have this kind of one-dimensional tuning — for example, it is easy to find V1 cells which are notably multidimensional (e.g. [Touryan et al., 2002, Rust et al., 2003]) — but it is not hard to see that our observations are consistent with and extend the classical "cosine" model of MI tuning [Georgopoulos et al., 1986, Moran and Schwartz, 1999].

We support the qualitative one-dimensional picture in Fig. 2.5 with

two somewhat more quantitative results. First, we could find no two-dimensional parametric model which fit the nonlinearity (in a likelihood sense) better than a simple one-dimensional model in any of the cells we examined (after suitable correction for differences in dimensionality [Schwarz, 1978]). Second, the mutual information in the second most modulatory axis $I(spike; < \vec{k}_2, \vec{x} >)$ is not significantly different from zero (according to a Monte Carlo test constructed by simulating spike trains from a one-dimensional model whose parameters were matched and whose inputs were identical to those of the real cell, then estimating K and $I(spike; < \vec{k}_2, \vec{x} >)$ for this model, repeating the procedure often enough to construct a nonparametric estimate of the null distribution to test against). Further details on $\vec{k}_1$, f, and this information analysis will be presented elsewhere [Paninski et al., 2002, Shoham et al., 2003, Paninski et al., 2003b].

2.7 Lower bounds

Our final mathematical results are lower bounds on the convergence rates of any possible K-estimator; these kinds of bounds provide rigorous measures of the difficulty of a given estimation problem, or of the efficiency of a given estimator. The first lower bound is local, in the sense that we assume that the true parameter is known *a priori* to be in some small neighborhood of parameter space. Recall that the Hellinger metric between any two densities is defined as (half of) the L_2 distance between the square roots of the

densities.

Theorem 29 (Local (Hellinger) lower bound). *For simplicity, let p be standard normal. For any fixed differentiable f, uniformly bounded away from 0 and 1 and with a uniformly bounded derivative f', and any Hellinger ball $\mathcal{F}$ around the true parameter (f, K),*

$$\liminf_{N\to\infty} N^{1/2} \inf_{\hat{K}} \sup_{\mathcal{F}} E(Error(\hat{K})) \geq \left(\sigma(p)(E_p(\frac{|f'|^2}{f(1-f)}))^{1/2}\right)^{-1} \sqrt{\dim\ X - 1}.$$

The second infimum above is taken over all possible estimators $\hat{K}$. The right-hand side plays the role of the inverse Fisher information in the Cramer-Rao bound and is derived using a similarly local analysis; see [Jongbloed, 2000] for details on the Hellinger technique, or [Gill and Levit, 1995] on the Bayesian Cramer-Rao technique.

Global bounds are more subtle. We want to prove something like:

$$\liminf_{N\to\infty} a_N \inf_{\hat{K}} \sup_{\mathcal{F}(\epsilon)} E(Error(\hat{K})) \geq C(\epsilon),$$

where $\mathcal{F}(\epsilon)$ is some large parameter set containing, say, all K and all f for which some relevant measure of tuning is greater than ϵ, a_N is the corresponding convergence rate, and $C(\epsilon)$ plays the role of $\alpha(\hat{K})$ from the previous sections. So far, our most interesting results in this direction are negative:

Theorem 30 (ϕ-divergences are poor indices of K-difficulty). *Let $\mathcal{F}(\epsilon)$ be the set of all (K, f) for which the ϕ-divergence "information" be-*

tween $\vec{x}$ and spike is greater than ϵ, that is,

$$D_\phi(p(K\vec{x}, spike);\ p(spike)p(K\vec{x})) > \epsilon.$$

Then, for $\epsilon > 0$ small enough, for any putative convergence rate a_N,

$$\liminf_{N\to\infty} a_N \inf_{\hat{K}} \sup_{\mathcal{F}(\epsilon)} E(Error(\hat{K})) = \infty.$$

In other words, strictly information-theoretic measures of tuning do not provide a useful index of the difficulty of the K-learning problem; the intuitive explanation of this result is that purely measure-theoretic distance functions, like ϕ-divergences, ignore the topological and vector space structure of the underlying probability measures, and it is exactly this structure that determines the convergence rates of any efficient K-estimator. To put it more simply, the learnability of K depends on the smoothness of f, just as we saw in the last section (c.f. Theorem 26), a common theme in non-parametric statistics.

2.8 Conclusion and directions for future work

We have presented here a fairly detailed analysis of the statistical properties of the LN model (2.1). In particular, we have attempted to elucidate when and why the common estimators for the LN model parameters work well, or not. More importantly, we have provided a new estimator which is guaranteed to recover the true parameters in much greater generality than was previously possible. We hope that our results will find application in

understanding the neural processing of naturalistic stimuli; as mentioned briefly in section (2.6.3), these methods have already led to a better understanding of the neural coding of dynamic hand movement signals in primary motor cortex.

We take this opportunity to outline one obvious avenue for future work: how do we extend the basic LN model (2.1) in a way that allows us to capture more of the details of the neural code, while at the same time retaining some of the simplicity that allows us to estimate the model? We discuss three such possible extensions below.

2.8.1 Non-Poisson effects

As noted in the introduction, model (2.1) generates spike trains which are (conditionally inhomogeneous) Poisson processes (note that, even if the stimulus ensemble is time-translation invariant, the spike train is not necessarily a marginally homogeneous Poisson process); given the input signal $\vec{x}$, the spikes in one time bin do not depend on those in any other nonoverlapping bin. We can extend this model by allowing spikes which are close to each other in time to be dependent (the importance of such an extension has been noted in several contexts; see, e.g, [Berry and Meister, 1998, Brown et al., 2002, Pillow and Simoncelli, 2003]). Some natural questions immediately arise. Does the standard spike-triggered analysis fail in this case? If so, why? Can we correct for these non-Poisson effects? We can give at least preliminary answers to all of these questions, at least in the

following special case:

$$p(spike|\vec{x}, s_-) = f(<\vec{k}_1, \vec{x}>, <\vec{k}_2, \vec{x}>, \ldots, <\vec{k}_m, \vec{x}>)g(T(s_-)). \quad (2.8)$$

Here T is some arbitrary statistic of s_-, the spike train up to the present time (e.g., T could encode the time since the last spike); the "modulation function" g maps the range of T into the half-interval $[0, \infty)$. The only conditions on f and g are those necessary to make $p(spike|\vec{x}, s_-)$ a regular conditional distribution (aside from measurability issues, it is sufficient that $f, g \geq 0,\ fg \leq 1\ \forall(K\vec{x}, s_-)$).

To see why the memory effects displayed by (2.8) complicate the analysis presented in the previous sections, recall the basic idea behind Chichilnisky's proof of the fact that, for model (2.1), whenever $p(\vec{x})$ is radially symmetric, $E(\hat{K}_{STA})$ lies in K (we are abusing notation slightly; K here denotes the subspace generated by K, which is assumed to be one-dimensional, as in section 2.3). We will write $E(\hat{K}_{STA})$ out and show the essential point of the proof; then we will show why the memory effects seen in (2.8) cause problems, and how these problems can be "fixed," in some suitable sense. We have

$$\begin{aligned} E(\hat{K}_{STA}) &= \int p(\vec{x}|spike)\vec{x}d\vec{x} \\ &= \int p(spike|\vec{x})\frac{p(\vec{x})}{p(spike)}\vec{x}d\vec{x} \\ &= \int f(<\vec{K}, \vec{x}>)\frac{p(\vec{x})}{p(spike)}\vec{x}d\vec{x}. \end{aligned}$$

The first equality is Bayes, the second (2.1). The essential point is that the

conditional probability of a spike given $\vec{x}$ depends only on $< \vec{K}, \vec{x} >$ - the proof that $E(\hat{K}_{STA}) \in K$ follows immediately (after a suitable change of basis). This key equality does not hold in general for (2.8):

$$\begin{aligned} p(spike|\vec{x}) &= \int p(spike|\vec{x}, s_-)p(s_-|\vec{x})ds_- \\ &= \int f(< \vec{K}, \vec{x} >)g(T(s_-))p(s_-|\vec{x})ds_- \\ &= f(< \vec{K}, \vec{x} >) \int g(T(s_-))p(s_-|\vec{x})ds_- \\ &= f(< \vec{K}, \vec{x} >)h(\vec{x}). \end{aligned}$$

The first equality is (2.8), the second linearity; the last is by way of definition: h is an abbreviation for the conditional expectation of $g(T(s_-))$ given $\vec{x}$. If $g \equiv 1$ (as in (2.1)), then $h(\vec{x}) \equiv 1$, and we recover $E(\hat{K}_{STA}) \in K$. However, in general, h is nonconstant in $\vec{x}$: h depends on $\vec{x}$ not only through its projection onto $\vec{K}$, but also through its projection on all time-translates of K to the left (i.e., all functions $k_{-\tau}$ such that $k_{-\tau}(t) = k(t + \tau)$, for some $k \in K$ and $\tau > 0$). Most K, of course, are not time-translation invariant. This breaks the proof and the result; indeed, it is easy to think of simple (non-pathological) examples of f, g, and radially symmetric $p(\vec{x})$ for which $E(\hat{K}_{STA}) \notin K$.

So we need to modify $\hat{K}_{STA}$ somehow to bring its expectation back into the desired subspace. Assume for simplicity that g is bounded below away from zero and that g and $T(s_-)$ are known (the simultaneous estimation of f, g, and K appears to be more difficult; no consistent estimator

for (f, g, K) seems to be known, although attempts have appeared, e.g., [Berry and Meister, 1998]). [Aguera y Arcas et al., 2001] suggest ignoring all spikes for which $g(T(s_-)) \neq 1$: i.e., form

$$\hat{K}_{STA^*} \equiv \frac{1}{N_s} \sum_{i \in S} \delta(g(T(s_{i-})) - 1)\vec{x}_i,$$

where S, again, indicates the set of stimuli corresponding to spikes, and δ the usual Dirac functional. However, the above string of equations shows that this procedure can actually make the situation worse: this effectively sets g equal to zero at all of these points where $g \neq 1$, which in many cases makes h more strongly $\vec{x}$-dependent, not less. In addition, of course, ignoring these "bad" spikes is expensive from a data collection point of view. An obvious alternative would be to form

$$\hat{K}_{STA_*} \equiv \frac{1}{N_s} \sum_{i \in S} g(T(s_{i-}))^{-1}\vec{x}_i.$$

It is easy to see, from the above discussion, that $E(\hat{K}_{STA_*}) \in K$.

More complete analysis of this kind of model and estimator would clearly be useful.

2.8.2 Integrate-and-fire models and logconvexity

Here we analyze an integrate-and-fire version of the LN idea[1]. More precisely, consider a model for which the (dimensionless) subthreshold voltage

[1]Joint work with Jonathan Pillow.

variable V evolves as

$$dV(t) = \left(\vec{k} * \vec{x}(t) - gV(t)\right)dt + N(t) - (1 - V_{reset})\delta(V - 1), \qquad (2.9)$$

where g denotes the membrane conductance, $V_{reset} < 1$ the reset potential, $\vec{k} * \vec{x}$ the convolution of the input signal $\vec{x}(t)$ with the (single) kernel $\vec{k}$, and N an unobserved (hidden) noise process. The "leak" and "threshold" potential here are set at 0 and 1, respectively; the cell emits a single spike each time $V = 1$, and the voltage decays back to 0 in the absence of input.

This model, while clearly not completely satisfying from a biophysical point of view, is at least a step away from the essentially phenomenological realm of the LN model towards something more mechanistic [Reich et al., 1998]. (It is not hard to show that this step is nontrivial in the sense that the integrate-and-fire model does not fit into the LN framework; more precisely, under weak conditions on the noise process N, the IF model does not have the factored form of expression (2.1), or even of (2.8), for any finite m.) See [Pillow and Simoncelli, 2003], [Paninski et al., 2003c], and http://www.cns.nyu.edu/∼liam/adapt.html for further analysis of a few of the interesting differences between integrate-and-fire and LN models; [Pillow and Simoncelli, 2003], in particular, discuss what is effectively a zero-noise limit of the estimation problem we consider below.

The estimation problem for this model is identical to that considered in our LN work: how do we learn $\vec{k}$ from samples of $\vec{x}$, together with the concurrent spike times? For the LN model, the difficulty is that we do not

know f, the nonlinearity following the filtering stage. Here, in contrast, the nonlinearity is completely determined by only a few parameters. The obvious approach for estimating k and the parameters of the nonlinearity is maximum likelihood; the usual asymptotic theory for the MLE applies (consistency, asymptotic normality, and the asymptotic bias and variance rates in the sample size N are all easily obtained via straightforward calculations using Fisher information, etc.). Thus the statistical problem here is handled completely by standard techniques.

The computational problem, on the other hand, is more interesting. To compute the MLE, we need to compute the likelihood and develop an algorithm for maximizing it. Our main contribution here is that this likelihood function is log-convex in the parameters if N is a Gaussian noise process, independent of the input (see [Pillow et al., 2003] for details on efficient likelihood computations). This logconvexity, in turn, implies that any ascent algorithm will converge to the MLE without getting trapped in any local maxima. Thus the optimization problem inherent in computing the MLE — and by extension, the construction of a computationally and statistically efficient estimator — is tractable.

The likelihood function for this model is easily computed. Basically, the likelihood is a product over spikes (by the obvious conditional renewal property of the model); each multiplicand, in turn, is a Gaussian integral over a certain set, namely, the probability that the Gaussian voltage process induced by the (known) stimulus $\vec{x}_i$ convolved with the (unknown) kernel

$\vec{k}$ and generated by the Brownian takes its value within the set C_i defined as follows:

$$C_i = \left\{ \left(V(s) : \ V(s) < 1, \ 0 < s < t_i \right) \ \cap \ \left(V(t_i) \geq 1 \right) \right\},$$

i.e., the set of voltage traces that stay subthreshold only until the time of the spike, t_i. Symbolically, then, define the likelihood as

$$p_{\{\vec{x}_i,t_i\}}(\vec{k}, g, \sigma, V_{reset}) = \prod_i \int_{C_i} G(\vec{x}_i, \vec{k}, \sigma, g, V_{reset}),$$

where G denotes the obvious Gaussian probability (see the appendix for details), and the product is over all spikes. This likelihood is the object we need to maximize as a function of the parameters $(\vec{k}, \sigma, g, V_{reset})$.

Our main result is the following. We say that a smooth function on some Euclidean domain has no local extrema if the set of points at which the gradient vanishes is connected and contains a global extremum; thus all "local extrema" are in fact global, if a global extremum exists. (The existence of a global maximum in our case is assured asymptotically by standard results on the MLE [van der Vaart, 1998].)

Theorem 31. *The likelihood $p_{\{\vec{x}_i,t_i\}}(\vec{k}, g, \sigma, V_{reset})$ has no local extrema in $(\vec{k}, g, \sigma, V_{reset})$, for any data $\{\vec{x}_i, t_i\}$.*

It is worth saying a few words about logconvexity and its relevance for the MLE. The classical approach for establishing the nonexistence of local extrema is convexity: convexity of a function on Euclidean space obviously

precludes any local extrema. However, the basic idea can be extended with the use of any invertible function: clearly, if f has no local extrema, neither will $g(f)$, for any strictly increasing real function g. The logarithm, of course, is a natural choice for g in any probabilistic context in which independence plays a role. So logconvexity buys us everything convexity does, for less effort (since sums, as we will see in the proof, are much easier to work with than products). Indeed, since convexity of a function f is strictly stronger than logconvexity (as is easily demonstrated using Jensen's inequality), logconvexity is often a powerful tool even in situations for which convexity is useless. This basic idea seems to be less well-known than it should be.

We should also note that the proof extends without difficulty to some other noise processes which generate logconcave densities (where white noise has the standard Gaussian density); for example, the proof is nearly identical if N is allowed to be colored Gaussian noise with nonzero drift.

2.8.3 Network effects

The final extension is perhaps the simplest; for this reason, it might turn out to be the most powerful. The basic idea is that the mathematics of the LN model don't care what the "inputs" $\vec{x}$ are; we already emphasized, for example, that $\vec{x}$ could include either sensory or motor data. Therefore we can also let $\vec{x}$ include neural data — either from different cells which might have been recorded simultaneously on a multielectrode array, or from the

same cell at positive leads or lags — without having to modify any of the analytical techniques discussed above. More precisely, we modify (2.1) to

$$p(spike|\vec{x}_0, \vec{n}) = f(< \vec{k}_0, \vec{x}_0 > + < \vec{k}_n, \vec{n} >),$$

where $\vec{n}$ denotes the side neural information (a possibly nonlinearly transformed vector of spike counts from nearby cells, or some vector representation of the recent firing history of the cell in question, or some combination thereof). The linear functional $\vec{k}_n$ here could be thought of as an appendage of the original kernel $\vec{k}_1$. Clearly, this model is still of LN form if we let $\vec{k}_1 = \vec{k}_0 \otimes \vec{k}_n$ and $\vec{x} = \vec{x}_0 \otimes \vec{n}$. While this augmented model appears at first sight to be a rather crude modification of (2.1), much of the rough intuition used to be justify the LN model also applies unchanged here (recall the brief discussion in the introduction).

Physical interpretation aside, the model is useful statistically: the inclusion of these side neural effects can increase the predictability of some MI neurons greatly, even given the full kinematic signal. We show an example in Fig. 2.6; for this cell, inclusion of the population data significantly increased the predictability of the spike train, as measured both by $I(spike; < \hat{k}_0, \vec{x} >)$ versus $I(spike; < \hat{k}_1, \vec{x} >)$ and by the peak observed conditional firing rate. We see a significant effect in the population summary data as well (Fig. 2.7, left). Note again that these nonlinearities and information values were calculated using cross-validation, so we are not simply observing overfitting to the extra parameters in $\vec{x}$.

Finally, it is worth noting that observing the network activity gives roughly the same amount of information as does observing the position of the hand (Fig. 2.7, right), despite the fact that we are only observing a tiny fraction of the full MI network (5-25 cells observed simultaneously for these plots). It is interesting to compare this result to that of [Tsodyks et al., 1999], who (implicitly) constructed a single-dimensional neural-only LN model using simple spike-triggered averaging (without the linear regression correction) for V1 cells, using voltage-sensitive dye images instead of multineuronal recordings.

We are currently examining the implications of our results for the problem of decoding this population neural activity into an estimate of the ongoing hand position signal, for applications to the design of neural prosthetics [Paninski et al., 2002, Shoham et al., 2003].

Appendix A: Proofs

A.1 $\hat{K}_{STA}$

Consistency of $\hat{K}_{RSTA}$: sufficiency. By the strong law of large numbers, the proof comes down to a bias calculation, and Chichilnisky's proof [Chichilnisky, 2001] for $\hat{K}_{STA}$ illustrates the source of this bias very nicely. First, the conditional expectation $E(< \vec{x}, \vec{k}_1 > | spike)$ exists by the finite-variance assumption on $p(\vec{x})$. Then, for $\hat{K}_{RSTA}$, we have the following string

of equalities:

$$\begin{aligned}
E(\hat{K}_{RSTA}) &= E(A\hat{K}_{STA}) \\
&= \int A\vec{x}p(\vec{x}|spike)d\vec{x} \\
&= \int A\vec{x}\frac{p(spike|\vec{x})}{p(spike)}p(\vec{x})d\vec{x} \\
&= \int A\vec{x}\frac{f(<\vec{K},\vec{x}>)}{p(spike)}p(\vec{x})d\vec{x} \\
&= \int A^{1/2}\vec{y}\frac{f(<\vec{K},A^{-1/2}\vec{y}>)}{p(spike)}p(A^{-1/2}\vec{y})|A|^{1/2}d\vec{y} \\
&= \int A^{1/2}\vec{y}\frac{f(<A^{-1/2}\vec{K},\vec{y}>)}{p(spike)}p(A^{-1/2}\vec{y})|A|^{1/2}d\vec{y} \\
&= A^{1/2}\int \vec{y}\frac{f(<A^{-1/2}\vec{K},\vec{y}>)}{p(spike)}p(A^{-1/2}\vec{y})|A|^{1/2}d\vec{y}.
\end{aligned}$$

The first two equalities are by definition, the third Bayes, the fourth (2.1), the fifth a linear change of coordinates $y = A^{1/2}x$, the sixth by the symmetry of $A^{-1/2}$, and the seventh by linearity. The rest of the proof follows [Chichilnisky, 2001] (see also section 2.8.1). □

Consistency of $\hat{K}_{STA}$: necessity. The claim is that if p is asymmetric, then there exists some f and $\vec{v}$ for which

$$\int \vec{x}\frac{f(<\vec{v},\vec{x}>)}{p(spike)}p(\vec{x})d\vec{x} \neq C_{f,\vec{v}}\vec{v},$$

for some scalar $C_{f,\vec{v}}$. The claim is equivalent to the following: if

$$\int \vec{x}\frac{f(<\vec{v},\vec{x}>)}{p(spike)}p(\vec{x})d\vec{x} = C_{f,\vec{v}}\vec{v} \qquad \forall f,\vec{v}, \tag{2.10}$$

then p is symmetric. It suffices to prove (2.10) for simple functions, that is, $f = 1$ on some set B, and $f = 0$ everywhere else. Thus condition (2.10) reduces to

$$\int_{<\vec{v},\vec{x}>\in B} <\vec{u},\vec{x}> p(\vec{x})d\vec{x} = 0 \qquad \forall B \in \mathcal{B},\ \vec{u}\perp\vec{v},$$

with $\mathcal{B}$ the class of all measurable sets. This, in turns, implies that the characteristic function (Fourier transform) of $p(\vec{x})$, $\check{p}(\vec{s})$, satisfies the following differential equation:

$$\frac{\partial\check{p}(\vec{s})}{\partial\vec{t}} = 0 \qquad \forall\vec{s}\perp\vec{t}.$$

Since $\check{p}$ is everywhere differentiable, by the finite-power assumption on p, the above equation implies that $\check{p}(\vec{s})$ is radially symmetric in $\vec{s}$, which, finally, implies the symmetry of p. □

Convergence rate $\alpha(\hat{K}_{RSTA})$. Assume p is elliptically symmetric. By the multivariate CLT and the computations above, $\hat{K}_{RSTA}$ is asymptotically normally distributed with mean

$$E(<\vec{x},\vec{k}_1>|spike)\vec{k}_1$$

and covariance matrix

$$\frac{1}{N_s}A^2C,$$

with

$$<\vec{v},C\vec{v}>=\int <\vec{v},\vec{x}>^2\, dp(\vec{x}|spike) \qquad \forall\vec{v}\in X'.$$

The asymptotic error behaves like the norm of this distribution orthogonal to $\vec{k}_1$, normalized by the projection of the mean of the distribution onto $\vec{k}_1$. The final result is

$$\alpha = \frac{(\text{trace } E^t A^2 C E)^{1/2}}{|E(< \vec{x}, \vec{k}_1 > |spike)|},$$

where E is any matrix whose columns are an orthonormal basis for the subspace of X' orthogonal to K. This reduces to the quoted result when p is Gaussian (in which case

$$E^t A^2 C E = E^t A E)$$

and white. □

A.2 $\hat{K}_{CORR}$

Consistency of $\hat{K}_{CORR}$: necessity. The argument for $\hat{K}_{CORR}$ is similar to that for $\hat{K}_{STA}$. We want to prove that if p is non-Gaussian, then there exists some f and $\vec{u} \perp \vec{v}$ for which

$$\int < \vec{u}, (\vec{x} - E_{p(\vec{x}|spike)}\vec{x}) >^2 \frac{f(< \vec{v}, \vec{x} >)}{p(spike)} dp(\vec{x}) \neq \int < \vec{u}, \vec{x} >^2 dp(\vec{x})$$

(recall we assumed that $E_{p(\vec{x})}\vec{x} = 0$). Without loss of generality, we assume that p is white, that is,

$$\int < \vec{u}, \vec{x} >^2 dp(\vec{x}) = 1 \qquad \forall \vec{u} : \ ||\vec{u}||_2 = 1;$$

translating into the contrapositive, again, we reformulate the claim as follows: if

$$\int < \vec{u}, (\vec{x} - E_{p(\vec{x}|spike)}\vec{x}) >^2 \frac{f(< \vec{v}, \vec{x} >)}{p(spike)} dp(\vec{x}) = 1 \qquad \forall f, \vec{u} \perp \vec{v}, \quad (2.11)$$

then p is Gaussian. The proof proceeds in two stages: first, we prove that the above condition implies that p is symmetric (making use of the result for $\hat{K}_{STA}$); then we prove that any symmetric p satisfying (2.11) is Gaussian. Again, we may restrict our attention to simple functions f.

First the symmetry. Note that (2.11) can be written as a mixture of conditional variances, given $< \vec{v}, \vec{x} >$. More formally, for simple f,

$$\begin{aligned} &\int < \vec{u}, (\vec{x} - E_{p(\vec{x}|spike)}\vec{x}) >^2 \frac{f(< \vec{v}, \vec{x} >)}{p(spike)} dp(\vec{x}) \\ &= \frac{1}{p(spike)} \int_{<\vec{v},\vec{x}>\in B} < \vec{u}, (\vec{x} - E_{p(\vec{x}|spike)}\vec{x}) >^2 \, dp(\vec{x}); \end{aligned}$$

in other words, (2.11) says that the conditional variance of $< \vec{u}, \vec{x} >$, given $< \vec{v}, \vec{x} > \in B$, is constant (for all vectors $\vec{u} \perp \vec{v}$ and all measurable sets B). Now consider disjoint subsets of B, B_1 and B_2, $B = B_1 \cup B_2$. It is clear that the following "mixture" equation holds:

$$\begin{aligned} p(< \vec{u}, \vec{x} > \quad &| < \vec{v}, \vec{x} > \in B) \\ &= \frac{1}{p(< \vec{v}, \vec{x} > \in B)} \Big(p(< \vec{v}, \vec{x} > \in B_1) p(< \vec{u}, \vec{x} > | < \vec{v}, \vec{x} > \in B_1) \\ &\quad + p(< \vec{v}, \vec{x} > \in B_2) p(< \vec{u}, \vec{x} > | < \vec{v}, \vec{x} > \in B_2) \Big). \end{aligned}$$

Now, since the mixture of two densities with the same positive, finite variance but different means has strictly greater variance than either of the two original densities, and each component in the above equation has the same

variance, each component must also have the same mean. That is,

$$\begin{aligned}\int_{<\vec{v},\vec{x}>\in B} <\vec{u},\vec{x}> p(\vec{x})d\vec{x} &= \int_{<\vec{v},\vec{x}>\in B_1} <\vec{u},\vec{x}> p(\vec{x})d\vec{x} \\ &= \int_{<\vec{v},\vec{x}>\in B_2} <\vec{u},\vec{x}> p(\vec{x})d\vec{x} \\ &= 0.\end{aligned}$$

The above equations hold for all such B, B_1, B_2, and are equivalent to condition (2.10), from the proof for $\hat{K}_{STA}$; thus p is symmetric.

Now, given that p is symmetric, we can write

$$p(\vec{x}) = g(||\vec{x}||_2^2),$$

for some scalar function g; it turns out that (2.11) provides us with a simple differential equation for p (and hence g) in Fourier space. For simple f and symmetric p, (2.11) reduces to

$$\int_{<\vec{v},\vec{x}>\in B} <\vec{u},\vec{x}>^2 dp(\vec{x}) = \int_{<\vec{v},\vec{x}>\in B} dp(\vec{x}) \qquad \forall B \in \mathcal{B}, \vec{u}\perp\vec{v}.$$

In the Fourier domain, this means that

$$\frac{\partial^2 \check{p}}{\partial \vec{t}^2} = -\check{p}(\vec{s}), \qquad \forall \vec{s}\perp\vec{t} : ||\vec{t}||_2 = 1.$$

Applying this equation to g, we find that

$$\frac{\partial \check{g}(s)}{\partial s} = -\check{g}/2,$$

i.e.,

$$\check{g}(s) = ce^{-s/2},$$

for some constant c; the proof is complete upon applying the inverse Fourier transform and normalizing. $\square$

Convergence rate $\alpha(\hat{K}_{CORR})$. Assume p is nondegenerate Gaussian. By the multivariate CLT, $\widehat{\Delta\sigma^2}$ is asymptotically normal with mean $\Delta\sigma^2$ and covariance

$$C \equiv E(\widehat{\Delta\sigma^2_{ij}} - \Delta\sigma^2_{ij})(\widehat{\Delta\sigma^2_{gh}} - \Delta\sigma^2_{gh}) = \frac{1}{N_s}(\sigma^2_{s,ih}\sigma^2_{s,jg} + \sigma^2_{s,ig}\sigma^2_{s,jh}),$$

where σ_s abbreviates the spike-triggered covariance matrix. Again, the proof relies on an analysis of the (normalized) behavior of the estimate on the orthogonal complement of K. This comes down to the usual local analysis, as follows.

Let eig_1 denote a top eigenvector map, that is, a map

$$\mathrm{eig}_1 : \Re^{(\dim X)^2} \to \Re^{\dim X},$$

taking a matrix to its (normalized) top eigenvector (in the case we will be dealing with, this map is uniquely defined almost surely). We know that, for N_s large enough,

$$\mathrm{eig}_1\widehat{\Delta\sigma^2} - \mathrm{eig}_1\Delta\sigma^2 \approx D\mathrm{eig}_1(\Delta\sigma^2)(\widehat{\Delta\sigma^2} - \Delta\sigma^2),$$

where $D\mathrm{eig}_1(\Delta\sigma^2)$ denotes the Jacobian matrix of eig_1 at the point $\Delta\sigma^2$; $\widehat{\Delta\sigma^2} - \Delta\sigma^2$ is normally distributed with mean zero and covariance C, and so we know everything we need to know about the asymptotic behavior of $\widehat{\Delta\sigma^2}$ if we can compute $D\mathrm{eig}_1(\Delta\sigma^2)$ on the (proper) subspace of $\Re^{(\dim X)^2}$ on

which C is a positive definite operator (this subspace is clearly contained in the space of all possible symmetric matrices, for example). The final result is that

$$\alpha(\hat{K}_{CORR})N_s^{-1/2} = \left(\text{trace } E(\sigma^2)^{-1}D\text{eig}_1(\Delta\sigma^2)CD\text{eig}_1(\Delta\sigma^2)^t(\sigma^2)^{-1}E^t\right)^{1/2}.$$

The computation of the derivative turns out to be fairly straightforward. We want to look at how much the symmetric perturbation ϵB, ϵ small, affects the i-th component of the first eigenvector of the symmetric matrix $A = VDV^t$, with V orthonormal and D diagonal. This is not difficult if V is the identity matrix; in this case, if D is zero everywhere but the first element λ, say, then a little direct computation shows that

$$\text{eig}_1(D + \epsilon B) - \text{eig}_1 D \approx \frac{\epsilon}{\lambda} Z_1 B,$$

where Z_1 is the operator mapping a matrix to its first column, after setting the first element to zero. The general result now follows after a change of basis or two:

$$\text{eig}_1(A + \epsilon B) - \text{eig}_1 A \approx \frac{\epsilon}{\lambda} V Z_1 V^t B V.$$

Plugging everything in, we get the stated result. □

A.3 $\hat{K}_\phi$

Consistency of $\hat{K}_\phi$. We want to prove that

$$\text{argmax}_V M_N(V) \to K$$

almost surely. According to arguments like those leading to Corollary 3.2.3 of [van der Vaart and Wellner, 1996], it suffices to prove the following two statements:

1) $M(V)$ has a well-separated, unique maximum at K;

2) $\sup_V |M_N(V) - M(V)| \to 0$ almost surely.

When we say that $M(K)$ is a "well-separated" maximum of $M(V)$, we mean that

$$M(K) > \sup_{V \in O^c} M(V),$$

where O is any open set containing K.

Part 1) is fairly straightforward. Under the conditions of the theorem, the sufficiency part of the data processing inequality ensures that K is a unique maximum. To see that this unique maximum is well-separated we need only note [van der Vaart and Wellner, 1996] that $M(V)$ is continuous in V under the conditions of the theorem, with compact domain, and that continuous functions on compact domains attain their suprema; thus, since $M(V)$ attains its maximum on the compact set O^c, $max_{V \in O^c} M(V)$ must be strictly less than the unique maximum $M(K)$.

Part 2) requires a little more effort. Letting $W_{a_N} * g$ denote the convolution of the kernel W_{a_N} with the function g, define the deterministic sequence of functions

$$M_N^*(V) \equiv \int_{spike,X} \frac{(W_{a_N} * p(spike, V\vec{x}))^2}{p(spike)(W_{a_N} * p(V\vec{x}))}.$$

Then the proof splits into two parts:

2a) $\sup_V |M_N(V) - M_N^*(V)| \to 0$ a.s.;

2b) $\sup_V |M_N^*(V) - M(V)| \to 0$.

We handle part a) with probability inequalities on uniform deviations of sample means from expectations (the standard VC inequalities [van der Vaart and Wellner, 1996, Devroye et al., 1996] are sufficient); since $M_N(V)$ is continuous on compact subsets of $\mathcal{G}_m(X)$ in the topology generated by uniform convergence and p is tight, the almost sure convergence follows.

We prove b) by noting that $M_N^*(V)$ is uniformly continuous in kernel width and V. Thus it is enough to prove pointwise convergence; this can be done under standard conditions on W [Devroye and Lugosi, 2001], either using Fourier transforms or by direct argument.

□

Bias of $\hat{K}_\phi$. We need to quantify the rate of decay of $M_N(V) - M(V)$. As indicated in the proof of the consistency theorem, this error has two parts: sample error and approximation error. The sample error, in addition, can be broken up into a bias term and a variance term. The bias term is what will cause us some problems, and it turns out that we can compute it explicitly.

We have

$$\begin{aligned} E(M_N(V) - M(V)) &= E\left(\int_{X,spike} \frac{\hat{p}(V\vec{x}, spike)^2}{\hat{p}(V\vec{x})\hat{p}(spike)} - \int_{X,spike} \frac{p(V\vec{x}, spike)^2}{p(V\vec{x})p(spike)} \right) \\ &= \int_X \left(E(\sum_{spike} (\frac{\hat{p}(V\vec{x}, spike)^2}{\hat{p}(V\vec{x})\hat{p}(spike)})) - \sum_{spike} (\frac{p(V\vec{x}, spike)^2}{p(V\vec{x})p(spike)}) \right), \end{aligned}$$

by definition, linearity, and Fubini.

Now we write out the expectation inside the integral. To simplify the computations, we assume either that we are dealing with the histogram estimator or that the kernel is a simple boxcar; this makes $\hat{p}$ a constant multiple of a binomial random variable. (The extension to more general kernels is not conceptually difficult [van der Vaart and Wellner, 1996], but precludes the direct calculations presented below.) Assume that N is large enough to replace $\hat{p}(spike)$ with $p(spike)$; this can be made rigorous with the usual exponential (Chernoff) inequalities [Devroye et al., 1996]. Let $W_a(x)$ denote the m-dimensional cube of width a centered on x, p^* the smoothed version of p, as above, s the event $(spike = 1)$, and $B(i, N, p)$ the

probability mass of a binomial with parameters N and p on count i; then

$$
\begin{aligned}
E \quad & (\sum_{spike}(\frac{\hat{p}(V\vec{x},spike)^2}{\hat{p}(V\vec{x})\hat{p}(spike)})) \approx E(\sum_{spike}(\frac{\hat{p}(V\vec{x},spike)^2}{\hat{p}(V\vec{x})p(spike)})) \\
= \quad & \sum_{i=0}^{N} B(i,N,\int_{W_a(V\vec{x})} p(V\vec{x}))\sum_{j=0}^{i} B(j,i,p^*(s|V\vec{x}))\frac{\frac{1}{p(s)}(\frac{j}{a^m N})^2 + \frac{1}{1-p(s)}(\frac{i-j}{a^m N})^2}{\frac{i}{a^m N}} \\
= \quad & \sum_{i=1}^{N} B(i,N,\int_{W_a(V\vec{x})} p(V\vec{x}))\frac{1}{ia^m N}\left[\frac{1}{p(s)}\left((ip^*(s|V\vec{x}))^2\right.\right. \\
& \left. + ip^*(s|V\vec{x})(1-p^*(s|V\vec{x}))\right) \\
& \left. + \frac{1}{1-p(s)}\left((i(1-p^*(s|V\vec{x})))^2 + ip^*(s|V\vec{x})(1-p^*(s|V\vec{x}))\right)\right] \\
= \quad & \frac{1}{a^m N}\left[\sum_{i=1}^{N} B(i,N,\int_{W_a(V\vec{x})} p(V\vec{x}))i\left(\frac{p^*(s|V\vec{x})^2}{p(s)} + \frac{(1-p^*(s|V\vec{x}))^2}{1-p(s)}\right)\right. \\
& \left. + p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))(\frac{1}{p(s)} + \frac{1}{1-p(s)})(1-(1-\int_{W_a(V\vec{x})} p(V\vec{x}))^N)\right] \\
= \quad & \sum_{spike}(\frac{p^*(V\vec{x},spike)^2}{p^*(V\vec{x})p(spike)}) + B_s(V),
\end{aligned}
$$

where we have abbreviated the sample bias in our estimate of $M(V)$ as

$$B_s(V) \equiv \frac{1}{a^m N}\int_X \left(\frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))}(1-(1-\int_{W_a(V\vec{x})} p(V\vec{x}))^N)\right).$$

To get a sense of how this behaves, let p be bounded and continuous, say; then the sample bias is roughly

$$\frac{1}{a^m N}\int_X \left(\frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))}(1-e^{-a^m N p(V\vec{x})})\right).$$

Now if p decays quickly enough (say it has compact support, to make things obvious), then the final term in the integral above tends to unity and we

are left with a bias in our estimate of $M(V)$ of order $(a^m N)^{-1}$. In turn, since the maximum of the above integral with respect to V is clearly not K in general, we are left with a bias of size up to $(a^m N)^{-1/2}$ in our estimate of $\text{argmax}_V M(V)$, as is easy to see after expanding $E(M_{\phi,N})$ about K. We should note that most of the above can be generalized to other choices of ϕ, using a second-order Taylor expansion.

If the sample bias for estimating $M(V)$ is of order $(a^m N)^{-1}$, and the approximation bias is of order a^r, say, for $r > 0$, then if we equate the two rates to minimize their sum we get that the optimal rate of decay in kernel width is

$$a \sim N^{\frac{-1}{r+m}},$$

corresponding to an optimal bias rate for $M(V)$ of

$$\text{bias} \sim N^{\frac{-r}{r+m}},$$

which in turn means that the optimal bias for estimating $\text{argmax}_V M(V)$ is of order $N^{\frac{-r}{2(r+m)}}$. □

Bias of the jackknifed kernel estimator. We write out the bias of the

jackknifed estimator, using the formula above:

$$
\begin{aligned}
ET_{JK} &= NET - \frac{N-1}{N}\sum_{i=1}^{N} ET_{-i} \\
&= N\frac{1}{a^m N}\int_X \frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))}(1-(1-\int_{W_a(V\vec{x})} p(V\vec{x}))^N) \\
&\quad -\frac{N-1}{N}\frac{N}{a^m(N-1)}\int_X \frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))}\Bigg(1 \\
&\quad -(1-\int_{W_a(V\vec{x})} p(V\vec{x}))^{N-1}\Bigg) \\
&= \frac{1}{a^m}\int_X \frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))}\Bigg((1-\int_{W_a(V\vec{x})} p(V\vec{x}))^{N-1} \\
&\quad -(1-\int_{W_a(V\vec{x})} p(V\vec{x}))^{N}\Bigg) \\
&= \frac{1}{a^m}\int_X \Bigg(\frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))} \\
&\quad (1-\int_{W_a(V\vec{x})} p(V\vec{x}))^{N-1}\int_{W_a(V\vec{x})} p(V\vec{x})\Bigg) \\
&= \int_X \Bigg(\frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))} \\
&\quad (1-\int_{W_a(V\vec{x})} p(V\vec{x}))^{N-1} W_a * p(V\vec{x})\Bigg).
\end{aligned}
$$

Under the conditions stated above, this dies exponentially. □

Convergence rates γ and α for $\hat{K}_\phi$. Representation (2.5) follows from a fairly classical first-order expansion; see, e.g., [Serfling, 1980] for background. $p_N m_V$ is the Frechet differential (aka the "functional derivative" to physicists) of $M_{\phi,N}(V)$ at $M_\phi(V)$ in the direction of $p_N - p$; the representation as a sum of i.i.d. random variables follows from the the linearity

of the differential. We obtain

$$m_V(\vec{x}, spike) = 2\frac{p(spike|V\vec{x})}{p(spike)} - \sum_{spike} \frac{p(spike|V\vec{x})^2}{p(spike)} - \int dp(\vec{x}) \left(\frac{p(spike|V\vec{x})}{p(spike)} \right)^2 .$$

The random variable $m_V(\vec{x}, spike)$ is bounded, thus obviously has finite variance. That the remainder term in (2.5) is $o_p(N^{-1/2})$ follows by computing its variance explicitly, roughly following the computation of the bias terms above. We skip the details. The convergence rate and limit distribution is obtained by applying Theorem 3.2.10 of [van der Vaart and Wellner, 1996] to $p_N m_V$. □

A.4 Lower bounds

Local (Cramer-Rao / Hellinger) lower bounds. The basic idea behind the proof is as follows. For any sufficiently regular finite-dimensional statistical model, the Cramer-Rao bound gives a lower bound on the convergence rate. The models we are dealing with are not finite-dimensional; nevertheless, we can apply the bounds to finite-dimensional submodels within the complete, infinite-dimensional family, and then try to make the bound as large as possible by choosing the most difficult submodel. By "most difficult" we mean, roughly: as close as possible to the true f, K, p in some probabilistic sense, but as far away as possible in the sense of the error metric (on the manifold $\mathcal{G}_m(X)$). In other words, we want the models to be as wrong as possible, but easily confusible with f, K, p.

The most obvious such submodel to try is obtained by keeping f fixed

(we assume p is fixed), and simply rotating K around in $\mathcal{G}_1(X)$. More concretely, we define our family to be

$$\mathcal{F}_0 \equiv \left\{ (q, g, V) : q = p, g = f, V \in \mathcal{G}_1(X) \right\}.$$

To apply Cramer-Rao to this family (under the stated conditions), we need to define an orthonormal basis of the tangent space to $\mathcal{G}_1(X)$ at K, $\{e_i\}_{1 \leq i < m}$; this induces a natural coordinate chart of $\mathcal{G}_1(X)$,

$$k_{\epsilon,i} \equiv (1 + \epsilon^2)^{-1/2}(k + \epsilon e_i).$$

The i-th component of the score vector when a spike occurs is given by

$$\frac{\partial log f_i}{\partial \epsilon} = \frac{f'(K\vec{x}) < e_i, \vec{x} >}{f(K\vec{x})};$$

plugging this into the asymptotic minimax form of the standard Cramer-Rao bound [Gill and Levit, 1995], we have

$$\liminf_{N \to \infty} N^{1/2} \inf_{\hat{K}} \sup_{\mathcal{F}} E(Error(\hat{K})) \geq (\text{trace } I_{\mathcal{F}_0}(p, f, K)^{-1})^{1/2},$$

where the Fisher information for $\mathcal{F}_0$ at the true model is given by

$$I_{\mathcal{F}_0}(p, f, K) = E_p\left(\frac{< e_i, \vec{x} >< e_j, \vec{x} > f'(< \vec{k}, \vec{x} >)^2}{f(< \vec{k}, \vec{x} >)(1 - f(< \vec{k}, \vec{x} >))} \right).$$

This reduces to the quoted result when p is, e.g., standard normal.

A more systematic approach to the search for "hard" subfamilies requires a more rigorous definition of the notion of "confusibility" between probability measures. While the detailed theory is beyond the scope of this paper, we mention that an appropriate measure of confusibility is given by

the Hellinger distance between two probability measures; recall that this distance is a kind of L_2 norm between (the square roots of) probability distributions, and can be written in our case as the square root of

$$H_p^2(f,K;h,V) \equiv \frac{1}{2}\int_X \left(f(K\vec{x})^{1/2} - h(V\vec{x})^{1/2} \right)^2 dp(\vec{x}).$$

For our purposes, it suffices to note that, for sufficiently close models (K,f) and (V,h),

$$H_p^2(f,K;h,V) \sim \int_X \left(\frac{(f(K\vec{x}) - h(V\vec{x}))^2}{f(K\vec{x})} \right) dp(\vec{x}).$$

Simple computations with this asymptotic form of Hellinger distance indicate a stronger subfamily:

$$\mathcal{F}_1 \equiv \left\{ (q,g,V) : q = p, V \in \mathcal{G}_1(X), g(t) = E_{p(\vec{x}|<V,\vec{x}>=t)} f(K\vec{x}) \right\}.$$

The final result is

$$\liminf_{N\to\infty} N^{1/2} \inf_{\hat{K}} \sup_{\mathcal{F}} E(Error(\hat{K})) \geq (\text{trace } I_{\mathcal{F}_1}(p,f,K)^{-1})^{1/2},$$

with

$$I_{\mathcal{F}_1}(p,f,K) = E_p \left(\frac{< e_i, \vec{y} >< e_j, \vec{y} > f'(< \vec{k}, \vec{x} >)^2}{f(< \vec{k}, \vec{x} >)(1 - f(< \vec{k}, \vec{x} >))} \right),$$

where we have made the abbreviation

$$\vec{y} = \vec{x} - E_{p(\vec{x}|<k,x>)} \vec{x}.$$

This inequality is in general stronger, but reduces to the first when p is, say, elliptically symmetric. □

Global minimax lower bound. We mimic [Ritov and Bickel, 1990] (Theorem 2). Let K_0 and K_N be separated by a distance of a_N. It suffices to put prior distributions π_N on the space of ϵ-tuned LN models supported on these two planes — that is, the conditional distributions given by model (2.1), with K given by K_0 or K_N, such that

$$D_\phi(p(K\vec{x}, spike);\ p(K\vec{x})p(spike)) > \epsilon$$

— such that the conditional error probability given N data samples of the best hypothesis test between K_0 and K_N converges to $1/2$ as $N \to \infty$. Since the best Bayesian test between two a_N-separated subspaces has error bounded away from zero, we have an order bound on the error of any minimax estimator, and the claim is proven. The basic idea behind the construction of the prior is to let "typical" functions (roughly, any function contained in the support of π_N) vary much more rapidly than the average distance between the projected samples $K\vec{x}_i$; this makes it impossible for any hypothesis test to discern the direction of the underlying conditional probability contour lines which run orthogonal to K. We skip the details, which are easy to verify given [Ritov and Bickel, 1990]. □

A.5 Logconvexity for the integrate-and-fire model

Proof. The proof is built on the following basic result (see, e.g., [Bogachev, 1998]).

Theorem ([Rinott, 1976]). *If p is a logconcave probability density func-*

tion on Euclidean space, then for any Borel sets A and B, for all t in $[0,1]$,

$$\log p(tA + (1-t)B) \geq t \log p(A) + (1-t) \log p(B),$$

that is, the corresponding measure is logconcave.

(For uniqueness of the global maximum, we would need the simple extension of this result that if p is strictly logconcave and A and B have positive p-measure, then the inequality is strict for all t in the open unit interval. To prove the nonexistence of local extrema, however, this is not necessary.)

Our proof basically consists of translating this result into the terminology of our problem. Let p be standard normal on Euclidean space (i.e., zero mean, identity covariance). Then p is obviously logconcave. Let C be any Borel set (e.g., the set satisfying the LIF constraints). Let $p_{\vec{x},\vec{k},\sigma,g,V_{reset}}$ denote the Gaussian probability density function induced by input $\vec{x}$ and parameters $(\vec{k}, \sigma, g, V_{reset})$. The mean, μ, of this Gaussian is equal to the solution of the noiseless version of the integrate-and-fire dynamics, on the interval $[0, t_i]$:

$$\frac{\partial V(t)}{\partial t} = < \vec{k}, \vec{x}_i(t) > -gV(t),$$

with initial data

$$V(0) = V_{reset},$$

that is:

$$V(t) = V_{reset} e^{-gt} + < \vec{k}, \vec{x} > * e^{-gt}.$$

The covariance Γ, in turn, is given by $E_g^t E_g$, where E_g is the convolution operator corresponding to e^{-gt}. Then

$$\begin{aligned} p_{\vec{x},\vec{k},\sigma,g,V_{reset}}(C) &= p(\Gamma^{-1/2}(C-\mu)) \\ &= p(\Gamma^{-1/2}(C - E_g(V_{reset}\delta(0) + < \vec{k}, \vec{x} >))) \\ &= p(\sigma^{-1}((E_g)^{-1}C - (V_{reset}\delta(0) + < \vec{k}, \vec{x} >))). \end{aligned}$$

Finally, for any fixed $\vec{x}, t$, $(\vec{k}_0, \sigma_0, g_0, V_{reset,0})$, $(\vec{k}_1, \sigma_1, g_1, V_{reset,1})$, and Borel C, define the sets

$$A = \sigma_0^{-1}\left((E_{g_0})^{-1}C - (V_{reset,0}\delta(0) + < \vec{k}_0, \vec{x} >)\right)$$

and

$$B = \sigma_1^{-1}\left((E_{g_1})^{-1}C - (V_{reset,1}\delta(0) + < \vec{k}_1, \vec{x} >)\right).$$

Our strategy now is to find some smooth, invertible parameterization of $(\vec{k}, g, \sigma, V_{reset})$ for which the convex translation of A into B corresponds to a convex line in the new parameter space; then we can use the theorem from Bogachev to say that the likelihood (i.e., the integral of the Gaussian process over the subthreshold spiking set C) given one piece of data is logconcave (for any data) in the new parameter space. The form of this parameterization should be clear enough: σ corresponds to a scale factor, $\vec{k}$ and V_{reset} translations, and g a rotation (E_g is unitary).

The rest is starightforward. Since the total loglikelihood function, given a collection of x and corresponding spike times, is a sum of logs of logconcave

functions (i.e., a sum of concave functions), the loglikelihood is concave and therefore has no local extrema, as the sets

$$\{x : f(x) >= y\}$$

are convex for a concave function f and any scalar level y. (Note that everything in sight is smooth, so our definition of local extrema in terms of vanishing gradients corresponds exactly to the more usual definition.) Now we only need to invert our parameterization of $(\vec{k}, g, \sigma, V_{reset})$; since diffeomorphisms can neither create nor destroy zero-gradient points, and the image of a convex set under a continuous map is connected, the proof is complete.

□

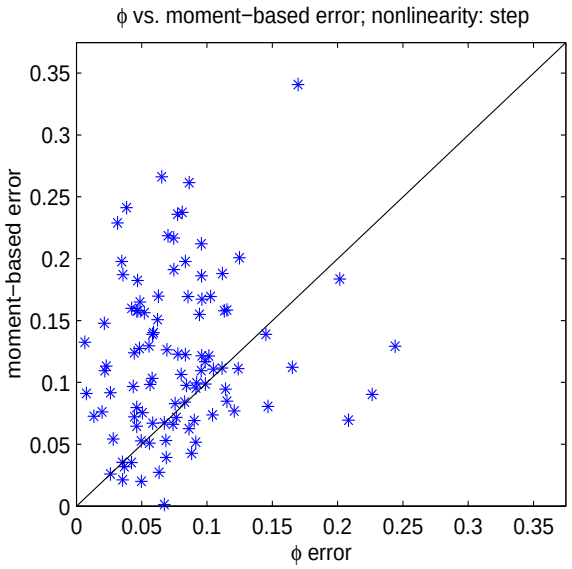

Figure 2.1: Plot of the error for $\hat{K}_\phi$ vs. that of $\hat{K}_{STA}$. $p(\vec{x})$ = Gaussian white noise; f is a step function, where the step position is chosen randomly. Axes index error in radian units. $N = 80$ and $\dim X = 3$ here; these small values were chosen for computational efficiency, but similar results are seen with larger values (see Fig. 2.3, for example). The error of $\hat{K}_\phi$ is slightly (but significantly) smaller than that of $\hat{K}_{STA}$ for these parameter settings.

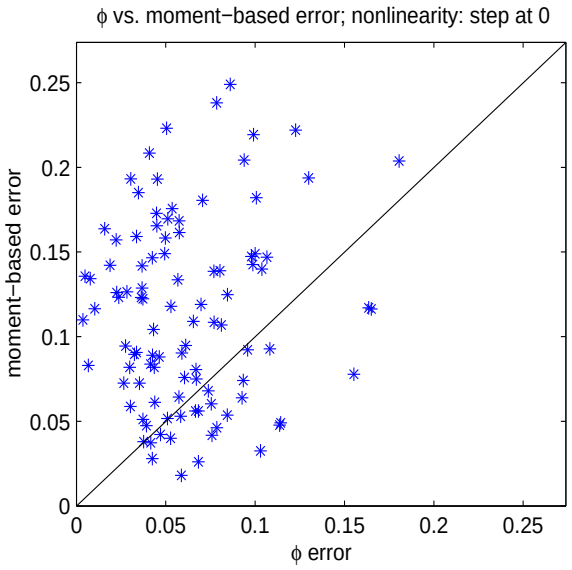

Figure 2.2: Plot of the error for $\hat{K}_\phi$ vs. that of $\hat{K}_{STA}$; parameters as in Fig. 2.1, except the step is always at zero. Conventions as in Fig. 2.1.

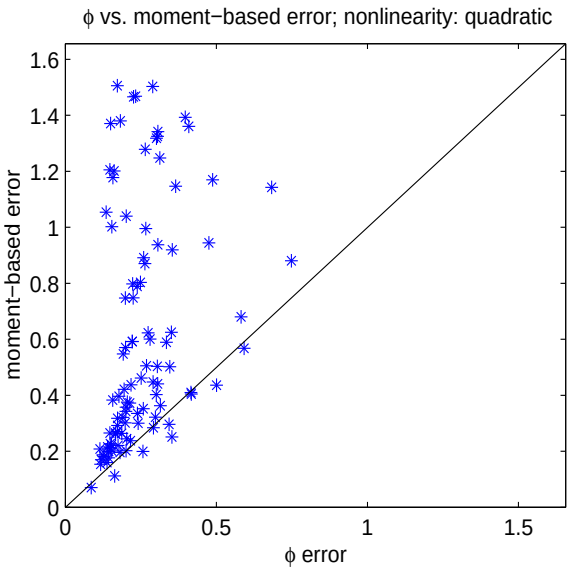

Figure 2.3: Plot of the error for $\hat{K}_\phi$ vs. that of $\hat{K}_{CORR}$. $p(\vec{x})$ = uniform on hypercube; $\vec{k}$ is chosen randomly; f is quadratic, with the center and scale chosen randomly. $N = 200$ and $\dim X = 10$ here; conventions as in Fig. 2.1.

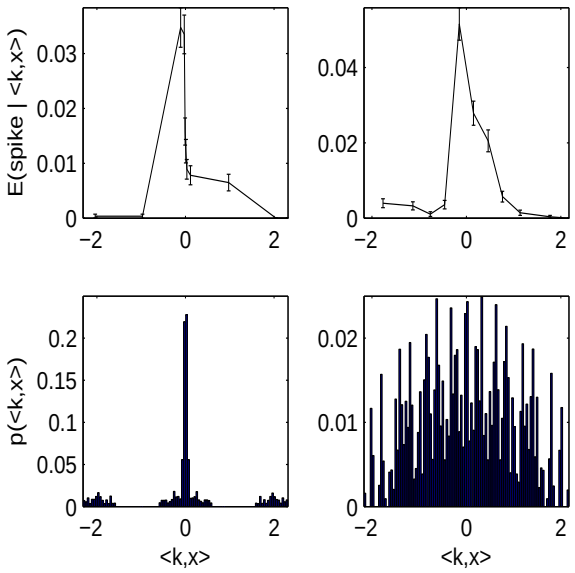

Figure 2.4: Comparison of second most informative axis in salamander retinal ganglion cell data, as estimated by $\hat{K}_{CORR}$ (left) and $\hat{K}_{\phi}$ (right). Top plots show nonlinearity (expected firing rate given $< \vec{k}, \vec{x} >$), estimated via adaptive histogram; bottom plots show raw marginal histograms $p(< \vec{k}, \vec{x} >)$. $\hat{K}_{\phi}$ extracts stronger tuning ($\approx 50\%$ greater peak firing rate; note difference in scales) by avoiding the artifact encountered by $\hat{K}_{CORR}$ (visible in left histogram).

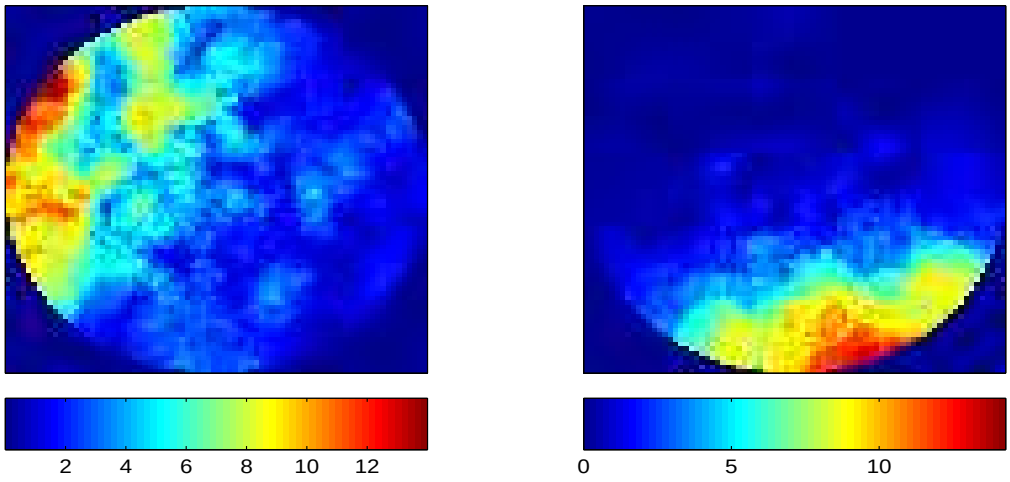

Figure 2.5: Example $\hat{f}(\hat{K}\vec{x})$ functions, computed from two different MI cells, with rank $\hat{K} = 2$; the x- and y-axes index $< \hat{k}_1, \vec{x} >$ and $< \hat{k}_2, \vec{x} >$, respectively, while the color axis indicates the value of $\hat{f}$ (the conditional firing rate given $\hat{K}\vec{x}$), in Hz. The scale on the x- and y-axes is arbitrary and has been omitted. $\hat{K}$ was computed using the ϕ-divergence estimator, and $\hat{f}$ was estimated using an adaptive kernel within the circular region shown (where sufficient data was available for reliable estimates). Note that the contours of these functions are approximately linear; that is, $\hat{f}(\hat{K}\vec{x}) \approx f_0(< \vec{k}_1, \vec{x} >)$, where $\vec{k}_1$ is the vector orthogonal to the contour lines and f_0 is a suitably chosen scalar function on the line.

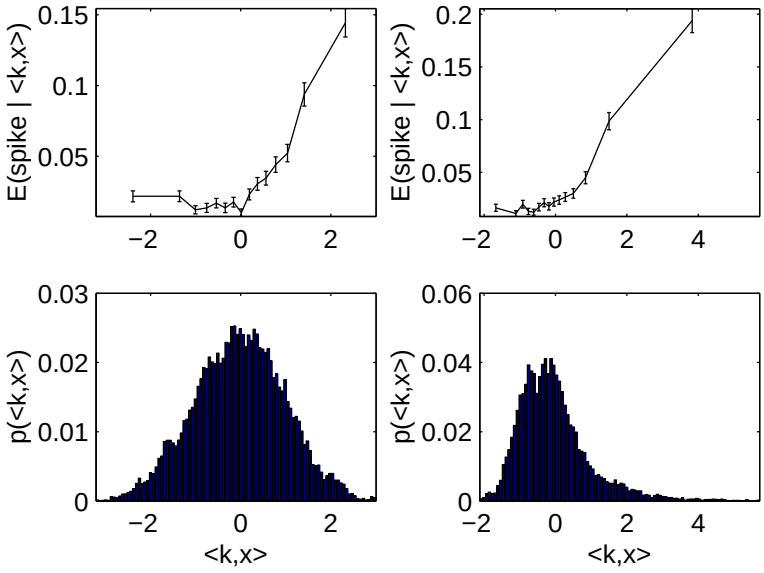

Figure 2.6: Comparison of estimated tuning given kinematic data only ($p(spike| < \hat{k}_0, \vec{x} >)$; left panels) versus kinematic data augmented with neural data recorded from adjacent electrodes ($p(spike| < \hat{k}_1, \vec{x} >)$; right). For this cell, network effects increased $I(spike| < \hat{k}_0, \vec{x} >)$ by approximately 50%, with a concurrent increase in observed peak conditional firing rate.

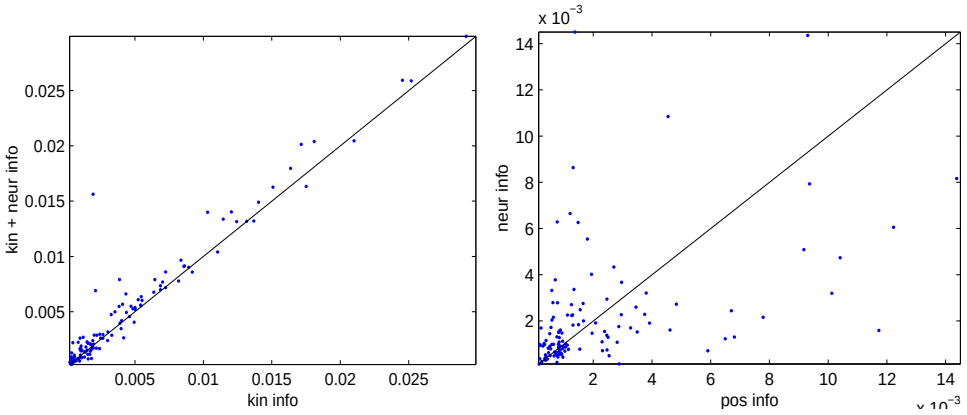

Figure 2.7: Population comparison of information values (bits, measured in 10 ms bins) for full model, including network effects, vs. model given kinematic data alone (left), and for neural model only vs. position only (right). Each point represents a single cell; diagonal line indicates unity.

CHAPTER 3

Information-theoretic design of experiments

3.1 Introduction

Many experiments are undertaken with the hope of elucidating some kind of "input-output" relationship: the experimenter presents some stimulus to the system under study and records the response. More generally, the experimenter places some observational apparatus in some state — for example, by pointing a microscope to a given location or selecting some subfield in a database stream — and records the subsequent observation. If the system is simple enough and enough observations are made, the resulting collection of data should provide a sufficiently precise description of the system's overall behavior.

Given this basic paradigm, in which the experimenter has some kind of control over what stimulus is chosen or what kind of data is collected, how do we design our experiments to be as efficient as possible? How can we learn

the most about the system under study in the least amount of time? This question becomes especially pressing in the context of high-dimensional, complex systems, where each input-output pair typically provides a small amount of information about the behavior of the system as a whole and opportunities to record responses are rare and/or expensive. In such cases, good experimental design can play an essential role in making the benefits of the experiment worth the cost.

How can we precisely define this intuitive concept of the "efficiency" of an experiment? First we have to define what exactly we mean by "experiment." We use the following simple model of experimental design here (we have neurophysiological experiments in mind, but our results are all completely general with respect to the identity of the system under study). The basic idea is that we have some set of models Θ, where each model θ indexes a given probabilistic input-output relationship. More precisely, a model is a set of probabilistic "input-output relationships" — regular conditional probability distributions $p(y|x,\theta)$ on Y, the set of possible output responses, given any input stimulus x in some space X. Therefore, if we know the identity of the model θ, we know the probability of observing any output y given any input x. Of course, we don't know θ precisely: our knowledge of the system is summarized in the form of a prior probability measure, $p_0(\theta)$, on Θ, and our goal is to reduce the uncertainty of this distribution as much as possible. To put everything together, the joint

probability of θ, x, and y is given by the following simple equation:

$$p(x, y, \theta) = p_0(\theta)p(x)p(y|\theta, x).$$

Now we can define the "design" of our experiment in a straightforward way: on any given trial, the design is specified completely by the choice of the input probability $p(x)$, the only piece of the above equation over which we have control. One common approach is to fix some $p(x)$ at the beginning of the experiment, then sample from this distribution in an i.i.d. manner for all subsequent trials, independently of which input-output pairs might have been observed on any previous trial. Alternatively, we could try to design our experiment — choose $p(x)$ — optimally in some sense, updating $p(x)$ on-line, on each trial, as more input-output data are collected and our understanding of the system increases. One natural idea would be to choose $p(x)$ in such a way that we learn as much as possible about the underlying model, on average. Information theory thus suggests we choose $p(x)$ to optimize the following objective function:

$$I(\{x, y\}; \theta) \tag{3.1}$$

where $I(.;.)$ denotes mutual information. In other words, we want to choose $p(x)$ adaptively, to maximize the information provided about θ by the pair $\{x, y\}$, given our current knowledge of the model as summarized in the posterior distribution given N samples of data:

$$p_N(\theta) = p(\theta|\{x_i, y_i\}_{1 \leq i \leq N}).$$

We will take this information-theoretic concept of efficiency as our starting point here. We note, however, that similar ideas have seen application in a wide and somewhat scattered literature: in statistics [Lindley, 1956], computer vision: [Denzler and Brown, 2000, Lee and Yu, 2001], machine learning [Luttrell, 1985, Axelrod et al., , Cohn et al., 1996, Freund et al., 1997, Mackay, 1992], conceptual psychology [Nelson and Movellan, 2000], psychophysics [Watson and Pelli, 1983, Pelli, 1987, Watson and Fitzhugh, 1990, Kontsevich and Tyler, 1999], medical applications [Parmigiani, 1998, Parmigiani and Berry, 1994], and neuroscience [Sahani, 1997]. These references all discuss, to a greater or lesser degree, the motivation behind various different design criteria, of which the information-theoretic criterion is well-motivated but certainly non-unique. For more general reviews of the theory of experimental design, see e.g. [Chaloner and Verdinelli, 1995] and [Fedorov, 1972]. In addition, several attempts have been made to devise algorithms to find the "optimal stimulus" of a neuron, where optimality is defined in terms of firing rate [Tzanakou et al., 1979, Foldiak, 2001, Nelken et al., 1994], but we should emphasize that the two concepts of optimality are not related in general, and turn out to be typically at odds (maximizing the firing rate of a cell does not maximize the amount we can expect to learn about the cell; see sections 3.3.1 and 3.3.2). Most recently, [Machens, 2002] proposed the maximization of the mutual information between the stimulus x and response y; again, though, this procedure does not directly maximize the amount of

information we gain about the underlying system θ.

Somewhat surprisingly, we have not seen any applications of the information-theoretic objective function (3.1) to the design of neurophysiological experiments (although see the abstract by [Mascaro and Bradley, 2002], who seem to have independently implemented the same idea in a simulation study). One major reason for this might be the computational demands of this kind of design (particularly for real-time applications), although these problems certainly do not appear to be intractable given modern computing power (see, e.g., [Kontsevich and Tyler, 1999] for a real-time application in which Θ is two-dimensional); we hope to address these important computational questions elsewhere.

The primary goal of this paper is to elucidate the asymptotic behavior of the *a posteriori* density p_N when we choose x according to the recipe outlined above; in particular, we want to compare the adaptive case to the more usual (i.i.d. x) case. Our main result (section 3.2) states that, under acceptably weak conditions on the models $p(y|x,\theta)$, the information-maximization strategy leads to consistent and efficient estimates of the true underlying model, in a natural sense. We also give a few simple examples to illustrate the applicability of our results 3.3, including a couple surprising negative examples that illustrate the nontriviality of our mathematical results (section 3.4).

3.2 Results

First, we note that the problem as posed in the introduction turns out to be slightly simpler than one might have expected, because $I(\{x,y\};\theta)$ is linear in $p(x)$:

$$\begin{aligned} I(\{x,y\}\ ;\ \theta) &= \int_X \int_Y \int_\Theta p(x,y,\theta) \log \frac{p(x,y,\theta)}{p(x,y)p_N(\theta)} \\ &= \int_X \int_Y \int_\Theta p(x,y,\theta) \log \frac{p(x)p_N(\theta)p(y|x,\theta)}{p(y|x)p(x)p_N(\theta)} \\ &= \int_X \int_Y \int_\Theta p(x,y,\theta) \log \frac{p(y|x,\theta)}{p(y|x)} \\ &= \int_X p(x) \int_Y \int_\Theta p_N(\theta)p(y|x,\theta) \log \frac{p(y|x,\theta)}{\int_\Theta p_N(\theta)p(y|x,\theta)}. \end{aligned}$$

This, in turn, implies that the optimal $p(x)$ must be degenerate, concentrated on the points x where I is maximal. Thus, instead of finding optimal distributions $p(x)$, we need only find optimal inputs x, in the sense of maximizing the conditional information between θ and y, given a single input x:

$$I(y;\theta|x) \equiv \int_Y \int_\Theta p_N(\theta)p(y|\theta,x) \log \frac{p(y|x,\theta)}{\int_\Theta p_N(\theta)p(y|x,\theta)}.$$

(We will assume throughout the paper that this function attains its supremum in X, and that some reasonable, though possibly non-deterministic, tie-breaking stategy exists when this maximum is not unique.)

Our main result is a "Bernstein-von Mises" - type theorem [van der Vaart, 1998]. The classical form of this kind of result says, basically, that if the posterior distributions are consistent (in the sense that

$p_N(U) \to 1$ for any neighborhood U of the true parameter θ_0) and the likelihood ratios are sufficiently smooth on average, then the posterior distributions $p_N(\theta)$ are asymptotic normal, with easily calculable asymptotic mean and variance. We adapt this result to the present case, where x is chosen according to the information-maximization recipe. It turns out that the hard part is proving consistency (c.f. section 3.4); we give the basic consistency lemma (interesting in its own right) first, from which the main theorem follows fairly easily.

Lemma 32 (Consistency). *Assume the following conditions:*

1. *The parameter space Θ is compact.*
2. *The loglikelihood $\log p(y|x,\theta)$ is Lipschitz in θ, uniformly in (x,θ), and y, with respect to some dominating measure.*
3. *The prior measure p_0 assigns positive measure to any neighborhood of θ_0.*
4. *The maximal divergence $\sup_x D_{KL}(\theta_0;\theta|x)$ is positive for all $\theta \neq \theta_0$.*

Then the posteriors are consistent: $p_N(U) \to 1$ in probability for any neighborhood U of θ_0.

Theorem 33 (Asymptotic normality). *Assume the conditions of Lemma 32, stengthened as follows:*

1. *Θ has a smooth, finite-dimensional manifold structure in a neighborhood of θ_0.*

2. *The loglikelihood $\log p(y|x,\theta)$ is continuously differentiable in θ, uniformly in (x,θ), and y, with respect to some dominating measure. Moreover, the Fisher information matrices*

$$I_\theta(x) = \int_Y \left(\frac{\dot{p}(y|x,\theta)}{p(y|x,\theta)}\right)^t \left(\frac{\dot{p}(y|x,\theta)}{p(y|x,\theta)}\right) p(y|\theta,x),$$

where the differential $\dot{p}$ is taken with respect to θ, are well-defined and continuous in θ, uniformly in x,θ in some neighborhood of θ_0.

3. *The prior measure p_0 is absolutely continuous in some neighborhood of θ_0, with a continuous positive density at θ_0.*

Then

$$||p_N - \mathcal{N}(\mu_N, \sigma_N^2)|| \to 0$$

in probability, where $||.||$ denotes variation distance and $\mathcal{N}(\mu_N, \sigma_N^2)$ denotes the normal density with mean μ_N and variance σ_N^2. Here

$$\sigma_N^2 = \left(\sum_{i=1}^N I_{\theta_0}(x)\right)^{-1},$$

and μ_N is asymptotically normal with mean θ_0 and variance σ_N^2.

Corollary 34. *If, in addition, the prior p_0 is absolutely continuous, with density bounded on the parameter space Θ, then the maximum a posteriori (MAP) estimator is consistent almost surely, with asymptotic distribution $\mathcal{N}(\theta_0, \sigma_N^2)$.*

Corollary 35. *If the determinant of the Fisher information matrices $I_{\theta_0}(x)$ has a unique maximum for some $I_{\theta_0}(x)'$, then*

$$N\sigma_N^2 \to (I_{\theta_0}(x)')^{-1}.$$

Thus, under these conditions, the information maximization strategy works, and works better than the i.i.d. x strategy (where the asymptotic variance σ^2 is inversely related to an average, not a maximum, over x, and is therefore generically larger).

A few words about the assumptions are in order. Most should be fairly self-explanatory: the conditions on the priors, as usual, are there to ensure that the prior becomes irrelevant in the face of sufficient posterior evidence; the smoothness assumptions on the likelihood permit the local expansion which is the source of asymptotic normality; and the condition on the maximal divergence function $\sup_x D_{KL}(\theta_0; \theta|x)$ ensures that distinct models θ_0 and θ are identifiable. Finally, some form of monotonicity or compactness on Θ is necessary here, mostly to bound the maximal divergence function $\sup_x D_{KL}(\theta_0; \theta|x)$ and its inverse away from zero (the lower bound, again, is to ensure identifiability; the necessity of the upper bound, on the other hand, will become clear in section 3.4); also, compactness is useful (though not necessary) for adapting certain Glivenko-Cantelli bounds [van der Vaart, 1998] for the consistency proof.

It should also be clear that we have not stated the results as generally as possible; we have chosen instead to use assumptions which are simple

to understand and verify, and to leave the technical generalizations to the interested reader. Our assumptions should be weak enough for most neurophysiological and psychophysical examples, for example, by assuming that parameters take values in bounded (though possibly large) sets and that tuning curves are not infinitely steep. The proofs of these three results are basically elaborations on Wald's consistency method and Le Cam's approach to the Bernstein-von Mises theorem [van der Vaart, 1998], and will be provided elsewhere.

3.3 Applications

3.3.1 Psychometric model

As noted in the introduction, psychophysicists have employed versions of the information-maximization procedure for some years [Watson and Pelli, 1983, Pelli, 1987, Watson and Fitzhugh, 1990, Kontsevich and Tyler, 1999]. References in [Watson and Fitzhugh, 1990], for example, go back four decades, and while these earlier investigators usually couched their discussion in terms of variance instead of entropy, the basic idea is the same. Our results above allow us to precisely quantify the effectiveness of this stategy (note, for example, that minimizing entropy is asymptotically equivalent to minimaizing variance, by our main theorem).

One general psychometric model is as follows. The response space Y is binary, corresponding to subjective "yes" or "no" detection responses. Let f

be "sigmoidal": a uniformly smooth, monotonically increasing function on the line, such that $f(0) = 1/2$, $\lim_{t\to-\infty} f(t) = 0$ and $\lim_{t\to\infty} f(t) = 1$ (this function represents the detection probability when the subject is presented with a stimulus of strength t). Let $f_{a,\theta} = f((t-\theta)/a)$; θ here serves as a location ("threshold") parameter, while a sets the scale (we assume a is known, for now, although of course this can be relaxed [Kontsevich and Tyler, 1999]). Finally, let $p(x)$ and $p_0(\theta)$ be some fixed sampling and prior distributions, respectively, both equivalent to Lebesgue measure on some interval Θ.

Now, for any fixed scale a, we want to compare the performance of the information-maximization strategy to that of the i.i.d. $p(x)$ procedure. We have by the corollary to theorem 33 that the most efficient estimator of θ is asymptotically unbiased with asymptotic variance

$$\sigma^2_{info} \approx (N \sup_x I_{\theta_0}(x))^{-1},$$

while the usual calculations show that the asymptotic variance of any efficient estimator based on i.i.d. samples from $p(x)$ is given by

$$\sigma^2_{iid} \approx (N \int_X dp(x) I_{\theta_0}(x))^{-1}.$$

The Fisher information is easily calculated here to be

$$I_\theta = \frac{(\dot{f}_{a,\theta})^2}{f_{a,\theta}(1-f_{a,\theta})}.$$

We can immediately derive two easy but important conclusions. First, there is just one function f^* satisying the assumptions stated above

for which the i.i.d. sampling strategy is as asymptotically efficient as information-maximization strategy; for all other f, information maximization is strictly more efficient. The extremal function f^* is the unique solution of the following differential equation:

$$\frac{df^*}{dt} = c\Big(f^*(t)(1 - f^*(t))\Big)^{1/2},$$

where the auxiliary constant $c = \sqrt{I_\theta}$ uniquely fixes the scale a. After some calculus, we obtain

$$f^*(t) = \frac{\sin(ct) + 1}{2}$$

on the interval $[-\pi/2c, \pi/2c]$ (and defined uniquely, by monotonicity, as 0 or 1 outside this interval). Since the support of the derivative of this function is compact, this result is not independent of the sampling density $p(x)$; if $p(x)$ places any of its mass outside of the interval $[-\pi/2c, \pi/2c]$, then σ^2_{iid} is always strictly greater than σ^2_{info}. This recapitulates a basic theme from the psychophysical literature comparing adaptive and nonadaptive techniques: when the scale of the nonlinearity f is either unknown or smaller than the scale of the i.i.d. sampling density $p(x)$, adaptive techniques are greatly preferable.

Second, a crude analysis shows that, as the scale of the nonlinearity $1/a$ shrinks, the ratio $\sigma^2_{iid}/\sigma^2_{info}$ grows approximately as a; this gives quantitative support to the intuition that the sharper the nonlinearity with respect to the scale of the sampling distribution $p(x)$, the more we can expect the information-maximization strategy to help.

3.3.2 Linear-nonlinear cascade model

We now consider a model that has received a growing amount of attention from the neurophysiology community (see, e.g., [Paninski, 2003a] for some analysis and relevant references). The model is of cascade form, with a linear stage followed by a nonlinear stage: the input space X is a compact subset of d-dimensional Euclidean space (take X to be the unit sphere, for concreteness), and the firing rate of the model cell, given input $\vec{x} \in X$, is given by the simple form

$$E(y|\vec{x}, \theta) = f(< \vec{\theta}, \vec{x} >).$$

Here the linear filter $\vec{\theta}$ is some unit vector in X', the dual space of X (thus, Θ is isomorphic to X), while the nonlinearity f is some nonconstant, nonnegative function on $[-1, 1]$. We assume that f is uniformly smooth, to satisfy the conditions of theorem 33; we also assume f is known, although, again, this can be relaxed. The response space Y — the space of possible spike counts, given the stimulus $\vec{x}$ — can be taken to be the nonnegative integers. For simplicity, let the conditional probabilities $p(y|\vec{x}, \theta)$ be parametrized uniquely by the mean firing rate $f(< \vec{\theta}, \vec{x} >)$; the most convenient model, as usual, is to assume that $p(y|\vec{x}, \theta)$ is Poisson with mean $f(< \vec{\theta}, \vec{x} >)$. Finally, we assume that the sampling density $p(x)$ is uniform on the unit sphere (this choice is natural for several reasons, mainly involving symmetry; see, e.g., [Chichilnisky, 2001, Paninski, 2003a]), and that the prior $p_0(\theta)$ is positive and continuous (and is therefore bounded away from

zero, by the compactness of Θ).

The Fisher information for this model is easily calculated as

$$I_\theta(x) = \frac{(f'(<\vec{\theta}, \vec{x}>))^2}{f(<\vec{\theta}, \vec{x}>)} P_{\vec{x},\theta},$$

where f' is the usual derivative of the real function f and $P_{\vec{x},\theta}$ is the projection operator corresponding to $\vec{x}$, restricted to the $(d-1)$-dimensional tangent space to the unit sphere at θ. The corollary to theorem 33 does not apply directly here, since $\det(I_\theta(x))$ is everywhere zero; nevertheless, using the symmetry in the problem, it is not hard to modify the argument to show that

$$\sigma^2_{info} \approx \left(N \max_{t \in [-1,1]} \frac{f'(t)^2 g(t)}{f(t)} \right)^{-1},$$

while

$$\sigma^2_{iid} \approx \left(N \int_{[-1,1]} dp(t) \frac{f'(t)^2 g(t)}{f(t)} \right)^{-1},$$

where $g(t) = \sqrt{1-t^2}$, $p(t)$ denotes the one-dimensional marginal measure induced on the interval by the uniform measure $p(x)$ on the unit sphere, and σ^2 in each of these two expressions multiplies the $(d-1)$-dimensional identity matrix.

Clearly, the arguments of subsection 3.3.1 apply here as well: the ratio $\sigma^2_{iid}/\sigma^2_{info}$ grows roughly linearly in the inverse of the scale of the nonlinearity. The more interesting asymptotics here, though, are in d. This is because the unit sphere has a measure concentration property [Milman and Schechtman, 1986, Talagrand, 1995]: as $d \to \infty$, the measure $p(t)$ becomes exponentially concentrated around 0. In fact, it is easy

to show directly that, in this limit, $p(t)$ converges in distribution to the normal measure with mean zero and variance d^{-2}. The most surprising implication of this result is seen for nonlinearities f such that $f'(0) = 0$, $f(0) > 0$; we have in mind, for example, symmetric nonlinearities like those often used to model complex cells in visual cortex. For these nonlinearities,

$$\frac{\sigma^2_{info}}{\sigma^2_{iid}} = O(d^{-2}) :$$

that is, the information maximization strategy becomes infinitely more efficient than the usual i.i.d. approach as the dimensionality of the spaces X and Θ grows.

3.4 Negative Examples

Our next two examples are more negative and perhaps more surprising: they show how the information-maximation strategy can fail, in a certain sense, if the conditions of the consistency lemma are not met. In each case, the method can be fixed using ad hoc methods; it is unclear at present whether a generally applicable modification of the basic information maximization strategy exists.

3.4.1 Two-threshold model

Let Θ be multidimensional, with coordinates which are "independent" in a certain sense, and the expected information obtained from one coordinate of

the parameter remains bounded strictly away from the expected information obtained from one of the other coordinates. Consider the following model.

$$p(1|x) = \begin{cases} .5 & -1 < x \leq \theta_{-1}, \\ f_{-1} & \theta_{-1} < x \leq 0, \\ .5 & 0 < x \leq \theta_1, \\ f_1 & \theta_1 < x \leq 1 \end{cases}$$

where $0 \leq f_{-1}, f_1 \leq 1$,

$$|f_{-1} - .5| > |f_1 - .5|,$$

are known and $-1 < \theta_{-1} < 0$ and $0 < \theta_1 < 1$ are the parameters we want to learn.

Let the initial prior be absolutely continuous with respect to Lebesgue measure; this implies that all posteriors will have the same property. Then, using the inverse cumulative probability transform and the fact that mutual information is invariant with respect to invertible mappings, it is easy to show that the maximal information we can obtain by sampling from the left is strictly greater than the maximal information obtainable from the right, uniformly in N. Thus the information-maximization strategy will sample from the left side forever, leading to a linear information growth rate (and easily-proven consistency) for the left parameter and non-convergence on the right. Compare the performance of the usual i.i.d. approach for choosing x (using any Lebesgue-dominating measure on the parameter space),

which leads to the standard root-N rate for both parameters (i.e., is strongly consistent in posterior probability).

Note that this kind of inconsistency problem does not occur in the case of sufficiently smooth $p(y|x,\theta)$, by our main theorem. However, the next example shows that the lack of consistency is not necessarily tied to the discontinuous nature of the conditional densities.

3.4.2 White noise models

We present two models of slightly different flavor; the basic mechanism of inconsistency is the same in each case. The samples x take values on the positive integers. The models live on the positive integers as well: θ is given by standard discrete 1) normal and 2) binary white noise process (that is, $p(\theta)$ is generated by an infinite sequence of standard normals and independent fair coins, respectively). The conditionals are defined as follows. For the first model, the observations y are Gaussian-contaminated versions of $\theta(x)$, that is, $y \sim \mathcal{N}(\theta(x), 1)$. For the second model, let y be drawn randomly from $q_{\theta(x)}$, where q_0 and q_1 are nonidentical measures on some arbitrary space.

Then it is not hard to show, for either model, that an experimenter using the information-maximization strategy will never sample from any x infinitely often. (For the second model, in fact, if the densities of q_0 and q_1 with respect to some dominating measure are unequal almost surely, then we will sample from each x just once, almost surely.) This again

implies a lack of consistency of the posterior (although, as above, we have a linear growth of information). The basic idea is that there will always be a more informative part of the sample space X to measure from, and the experimenter will never spend enough time in one place x to sufficiently characterize $\theta(x)$.

As in the last section, the standard i.i.d. approach (using any measure which does not assign zero mass to any of the integers) is consistent here. Note that, in contrast with the last example, the smoothness of the conditionals $p(y|x,\theta)$ (in the Gaussian model) does not rescue consistency. Nor is the inconsistency due to some pathology of differential entropy (the measures q_i can be discrete, even binary).

Conclusion

Quantitative, systems-level neuroscience — by which we mean the subfield most directly concerned with this neural coding problem — has developed rapidly over the past couple decades, nourished in part by conceptual and technical advances and in part by sustained growth of computing power and funding opportunities. This explosion in neuroscience research has, in turn, nourished a vigorous development of statistical methods for collecting, analyzing, and modeling complex, high-dimensional neural data. This thesis, we hope, adds to this literature. As such, the methods we present were developed mainly with spike train data from extracellular recordings in mind, but it should be clear that all of the techniques we have developed here can be applied generally, without any neural context at all.

It is also worth noting that all the work presented here turns out to have some information-theoretic flavor, although this unifying thread was not imposed *a priori*. One could take this as another instance of the "rightness" of information theory for questions in neural coding; while this statement contains at least a kernel of truth — for example, we leaned rather heavily

on the data processing inequality (chapters 1 and 2) and the source coding interpretation of mutual information (chapter 3) — it is just as likely that this merely reflects the author's point of view.

Our overarching goal in all of this was to put these statistical methods, and by extension any techniques that employ similar ideas as a "front end," on as firm a theoretical foundation as possible. Mathematical rigor has two important consequences here. First, we obtain a much clearer picture of when the methods can be expected to work and when they can be expected to fail. When we deal with high-dimensional, complex data, which typically cannot be viewed directly in any meaningful way, this kind of mathematical control on the confidence we can place in our results is essential, in the same way that error bars and significance levels are necessary ingredients in the interpretation of the results of classical statistical analyses. Second, we obtain a clearer picture of *why* the techniques work (or don't). The practical consequence, as we saw, is that we can systematically fix flaws we find in the methods, to design new techniques with improved performance.

Some brief philosophical discussion might be in order here, as our goals may seem somewhat abstruse from a physiological point of view. After all, it is often considered bad form to spend more time discussing methods than results, as we do here. However, we feel that at this stage of the development of systems neuroscience, where the physiological methods are relatively quite mature (we can simultaneously image and record from many neurons in almost any brain area of almost any kind of reasonably-

sized animal, while presenting almost any stimulus and recording almost any behavior), the development of reliable, powerful statistical techniques for understanding this complex neural data is of utmost importance. We view these statistical methods as a kind of technology, much like, say, the tungsten electrode or calibrated monitor. It is easy to forget how fundamentally our scientific views depend on available technology; as we emphasized above, developments in statistical thinking have changed the neuroscience community's worldview more than once, and we are confident that statistical techniques will continue to have a deep influence on our understanding of the nervous system.

BIBLIOGRAPHY

[Adrian, 1926] Adrian, E. (1926). The impulses produced by sensory nerve endings. *Journal of Physiology*, 61:49–72.

[Aguera y Arcas et al., 2001] Aguera y Arcas, B., Fairhall, A., and Bialek, W. (2001). What can a single neuron compute? *NIPS*, 13:75–81.

[Antos and Kontoyiannis, 2001] Antos, A. and Kontoyiannis, I. (2001). Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms*, 19:163–193.

[Axelrod et al.,] Axelrod, S., Fine, S., Gilad-Bachrach, R., Mendelson, S., and Tishby, N. The information of observations and application for active learning with uncertainty.

[Basharin, 1959] Basharin, G. (1959). On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability and its Applications*, 4:333–336.

[Begun et al., 1983] Begun, J., Hall, W., Huang, W., and Wellner,

J. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Annals of Statistics*, 11:432–452.

[Beirlant et al., 1997] Beirlant, J., Dudewicz, E., Gyorfi, L., and van der Meulen, E. (1997). Nonparametric entropy estimation: an overview. *International Journal of the Mathematical Statistics Sciences*, 6:17–39.

[Berry and Meister, 1998] Berry, M. and Meister, M. (1998). Refractoriness and neural precision. *Journal of Neuroscience*, 18:2200–2211.

[Bialek, 1987] Bialek, W. (1987). Physical limits to sensation and perception. *Annual Review of Biophysics and Biomolecular Structure*, 16:455–478.

[Bialek et al., 1991] Bialek, W., Rieke, F., de Ruyter van Steveninck, R., and Warland, D. (1991). Reading a neural code. *Science*, 252:1854–1857.

[Billingsley, 1965] Billingsley, P. (1965). *Ergodic theory and information*. Wiley, New York.

[Bogachev, 1998] Bogachev, V. (1998). *Gaussian Measures*. AMS, New York.

[Brenner et al., 2001] Brenner, N., Bialek, W., and de Ruyter van Steveninck, R. (2001). Adaptive rescaling optimizes information transmission. *Neuron*, 26:695–702.

[Brown et al., 2002] Brown, E., Barbieri, R., Ventura, V., Kass, R., and Frank, L. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural Computation*, 14:325–346.

[Brown et al., 1998] Brown, E., Frank, L., Tang, D., Quirk, M., and Wilson, M. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18:7411–7425.

[Buracas et al., 1998] Buracas, G., Zador, A., DeWeese, M., and Albright, T. (1998). Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron*, 5:959–969.

[Bussgang, 1952] Bussgang, J. (1952). Crosscorrelation functions of amplitude-distorted gaussian signals. *RLE Technical Reports*, 216.

[Carlton, 1969] Carlton, A. (1969). On the bias of information estimates. *Psychological Bulletin*, 71:108–109.

[Chaloner and Verdinelli, 1995] Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, 10:273–304.

[Chander and Chichilnisky, 2001] Chander, D. and Chichilnisky, E. (2001). Adaptation to temporal contrast in primate and salamander retina. *Journal of Neuroscience*, 21:9904–16.

[Chichilnisky, 2001] Chichilnisky, E. (2001). A simple white noise analysis

of neuronal light responses. *Network: Computation in Neural Systems*, 12:199–213.

[Chow and Teicher, 1997] Chow, Y. and Teicher, H. (1997). *Probability theory*. Springer, New York.

[Cohn et al., 1996] Cohn, D., Ghahramani, Z., and Jordan, M. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145.

[Cover and Thomas, 1991] Cover, T. and Thomas, J. (1991). *Elements of information theory*. Wiley, New York.

[Csiszar, 1967] Csiszar, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–317.

[Cucker and Smale, 2002] Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletins of the American Mathematical Society*, 39:1–49.

[Darbellay and Vajda, 1999] Darbellay, G. and Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45:1315–1321.

[Darling, 1953] Darling, D. (1953). On a class of problems related to the

random division of an interval. *Annals of Mathematical Statistics*, 24:239–253.

[Dayan and Abbott, 2001] Dayan, P. and Abbott, L. (2001). *Theoretical Neuroscience*. MIT Press.

[de Ruyter and Bialek, 1988] de Ruyter, R. and Bialek, W. (1988). Real-time performance of a movement-senstivive neuron in the blowfly visual system: coding and information transmission in short spike sequences. *Proc. R. Soc. Lond. B*, 234:379–414.

[Dembo and Zeitouni, 1993] Dembo, A. and Zeitouni, O. (1993). *Large deviations techniques and applications*. Springer, New York.

[Denzler and Brown, 2000] Denzler, J. and Brown, C. (2000). Optimal selection of camera parameters for state estimation of static systems: An information theoretic approach. *U. Rochester Technical Reports*, 732.

[Devore and Lorentz, 1993] Devore, R. and Lorentz, G. (1993). *Constructive approximation*. Springer, New York.

[Devroye et al., 1996] Devroye, L., Gyorfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer-Verlag, New York.

[Devroye and Lugosi, 2001] Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York.

[Diaconis and Freedman, 1984] Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12:793–815.

[Ditzian and Totik, 1987] Ditzian, Z. and Totik, V. (1987). *Moduli of smoothness*. Springer-Verlag, Berlin.

[Donoho and Liu, 1991] Donoho, D. and Liu, R. (1991). Geometrizing rates of convergence. *Annals of Statistics*, 19:633–701.

[Efron and Stein, 1981] Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Annals of Statistics*, 9:586–596.

[Everson and Roberts, 2000] Everson, R. and Roberts, S. (2000). Inferring the eigenvalues of covariance matrices from limited, noisy data. *IEEE Transactions on Signal Proceedings*, 48:2083–2091.

[Fedorov, 1972] Fedorov (1972). *Theory of Optimal Experiments*. Academic Press, New York.

[Fellows et al., 2001] Fellows, M., Paninski, L., Hatsopoulos, N., and Donoghue, J. (2001). Diverse spatial and temporal features of velocity and position tuning in mi neurons during continuous tracking. *SFN abstracts*, page 940.1.

[Foldiak, 2001] Foldiak, P. (2001). Stimulus optimisation in primary visual cortex. *Neurocomputing*, 38–40:1217–1222.

[Freund et al., 1997] Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168.

[Gedeon et al., 2003] Gedeon, T., Parker, A., and Dimitrov, A. (2003). Information distortion and neural coding. *Canadian Applied Math Quarterly*, Under review.

[Georgopoulos et al., 1986] Georgopoulos, A., Caminiti, R., and Kalaska, J. (1986). Neuronal population coding of movement direction. *Science*, 233:1416–1419.

[Gibbs and Su, 2002] Gibbs, A. and Su, F. (2002). On choosing and bounding probability metrics. *International Statistical Review*, 70:419–436.

[Gill and Levit, 1995] Gill, R. and Levit, B. (1995). Applications of the van trees inequality: A bayesian cramer-rao bound. *Bernoulli*, 1/2:59–79.

[Grenander, 1981] Grenander, U. (1981). *Abstract inference*. Wiley, New York.

[Hecht et al., 1942] Hecht, S., Shlaer, S., and Pirenne, M. (1942). Energy, quanta, and vision. *Jouranl of General Physiology*, 25:819–840.

[Hunter and Korenberg, 1986] Hunter, I. and Korenberg, M. (1986). The identification of nonlinear biological systems: Wiener and hammerstein cascade models. *Biological Cybernetics*, 55:135–144.

[Hyvarinen et al., 2001] Hyvarinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley, Boston.

[Johnstone, 2000] Johnstone, I. (2000). On the distribution of the largest principal component. Technical Report 2000-27, Stanford.

[Jongbloed, 2000] Jongbloed, G. (2000). Minimax lower bounds and moduli of continuity. *Statistics and Probability Letters*, 50:279–284.

[Knill and Richards, 1996] Knill, D. and Richards, W., editors (1996). *Perception as Bayesian Inference*. Cambridge University Press.

[Kolmogorov, 1993] Kolmogorov, A. (1993). *Information theory and the theory of algorithms*. Kluwer, Boston.

[Kontoyiannis, 1997] Kontoyiannis, I. (1997). Second-order noiseless source coding theorems. *IEEE Transactions Information Theory*, 43:1339–1341.

[Kontsevich and Tyler, 1999] Kontsevich, L. and Tyler, C. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*.

[Lee and Yu, 2001] Lee, T. and Yu, S. (2001). An information-theoretic framework for understanding saccadic behaviors. *NIPS*, 12.

[Lindley, 1956] Lindley, D. (1956). On a measure of information provided by an experiment. *Annals of Mathematical Statistics*, 29:986–1005.

[Loizou, 1998] Loizou, P. (1998). Mimicking the human ear: an introduction to cochlear implants. *IEEE Signal Processing Magazine*, 15:101–130.

[Luttrell, 1985] Luttrell, S. (1985). The use of transinformation in the design of data sampling schemes for inverse problems. *Inverse Problems*, 1:199–218.

[Machens, 2002] Machens, C. (2002). Adaptive sampling by information maximization. *Physical Review Letters*, 88:228104–228107.

[Mackay, 1992] Mackay, D. (1992). Information-based objective functions for active data selection. *Neural Computation*, 4:589–603.

[Marmarelis and Marmarelis, 1978] Marmarelis, P. and Marmarelis, V. (1978). *Analysis of physiological systems: the white-noise approach*. Plenum Press, New York.

[Mascaro and Bradley, 2002] Mascaro, M. and Bradley, D. (2002). Optimized neuronal tuning algorithm for multichannel recording. Unpublished abstract at http://www.compscipreprints.com/.

[McDiarmid, 1989] McDiarmid, C. (1989). *Surveys in Combinatorics*, chapter On the method of bounded differences, pages 148–188. Cambridge University Press.

[Miller, 1955] Miller, G. (1955). Note on the bias of information estimates. In *Information theory in psychology II-B*, pages 95–100.

[Milman and Schechtman, 1986] Milman, V. and Schechtman, G. (1986). *Asymptotic Theory of Finite Dimensional Normed Spaces*, volume 1200 of *Lecture Notes in Math.* Springer-Verlag.

[Moran and Schwartz, 1999] Moran, D. and Schwartz, A. (1999). Motor cortical representation of speed and direction during reaching. *Journal of Neurophysiology*, 82:2676–2692.

[Nelken et al., 1994] Nelken, I., Prut, Y., Vaadia, E., and Abeles, M. (1994). In search of the best stimulus: an optimization procedure for finding efficient stimuli in the cat auditory cortex. *Hearing Research*, 72:237–253.

[Nelson and Movellan, 2000] Nelson, J. and Movellan, J. (2000). Active inference in concept learning. *NIPS*.

[Nemenman et al., 2002] Nemenman, I., Shafee, F., and Bialek, W. (2002). Entropy and inference, revisited. *Advances in neural information processing*, 14.

[Paninski, 2003a] Paninski, L. (2003a). Convergence properties of some spike-triggered analysis techniques. *Network: Computation in Neural Systems*, Under revision.

[Paninski, 2003b] Paninski, L. (2003b). Estimation of entropy and mutual information. *Neural Computation*, In press.

[Paninski et al., 1999] Paninski, L., Fellows, M., Hatsopoulos, N., and Donoghue, J. (1999). Coding dynamic variables in populations of motor cortex neurons. *Society for Neuroscience Abstracts*, 25:665.9.

[Paninski et al., 2003a] Paninski, L., Fellows, M., Hatsopoulos, N., and Donoghue, J. (2003a). Spatiotemporal tuning properties for hand position and velocity in motor cortical neurons. *Journal of Neurophysiology*, Submitted.

[Paninski et al., 2002] Paninski, L., Fellows, M., Shoham, S., and Donoghue, J. (2002). Nonlinear encoding and decoding in primary motor cortex (mi). *Society for Neuroscience Abstracts*, 28.

[Paninski et al., 2003b] Paninski, L., Fellows, M., Shoham, S., Hatsopoulos, N., and Donoghue, J. (2003b). Nonlinear population models for the encoding of dynamic hand position signals in mi. *Computational Neuroscience Meeting*, Submitted.

[Paninski et al., 2003c] Paninski, L., Lau, B., and Reyes, A. (2003c). Noise-driven adaptation: in vitro and mathematical analysis. *Neurocomputing*, In press.

[Panzeri et al., 1999] Panzeri, S. Treves, A., Schultz, S., and Rolls, E. (1999). On decoding the responses of a population of neurons from short time windows. *Neural Computation*, 11:1553–1577.

[Panzeri and Treves, 1996] Panzeri, S. and Treves, A. (1996). Analytical

estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7:87–107.

[Parmigiani, 1998] Parmigiani, G. (1998). Designing observation times for interval censored data. *Sankhya A*, 60:446–458.

[Parmigiani and Berry, 1994] Parmigiani, G. and Berry, D. (1994). *Aspects of Uncertainty: A tribute to D. V. Lindley*, chapter Applications of Lindley Information Measure to the Design of Clinical Experiments, pages 333–352. Wiley, NY.

[Pelli, 1987] Pelli, D. (1987). The ideal psychometric procedure. *Investigative Ophthalmology and Visual Science (Supplement)*, 28:366.

[Perkel et al., 1967] Perkel, D., Gerstein, G., and Moore, G. (1967). Neuronal spike trains and stochastic point processes. *Biophysical Journal*, 7:391–440.

[Pillow et al., 2003] Pillow, J., Paninski, L., and Simoncelli, E. (2003). Log-convexity and the leaky integrate-and-fire model. *Computational Neuroscience Meeting*, submitted.

[Pillow and Simoncelli, 2003] Pillow, J. and Simoncelli, E. (2003). Biases in white noise analysis due to non-poisson spike generation. *Neurocomputing*, in press.

[Prakasa Rao, 2001] Prakasa Rao, B. (2001). Cramer-rao type integral inequalities for general loss functions. *TEST*, 10:105–120.

[Press et al., 1992] Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical recipes in C*. Cambridge University Press.

[Reich et al., 1998] Reich, D., Victor, J., and Knight, B. (1998). The power ratio and the interval map: Spiking models and extracellular recordings. *The Journal of Neuroscience*, 18:10090–10104.

[Reid and Shapley, 2002] Reid, R. and Shapley, R. (2002). Space and time maps of cone photoreceptor signals in macaque lateral geniculate nucleus. *Journal of Neuroscience*, 22:6158–6175.

[Rieke et al., 1997] Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1997). *Spikes: exploring the neural code*. MIT Press, Cambridge.

[Ringach et al., 1997] Ringach, D., G., S., and Shapley, R. (1997). A subspace reverse correlation technique for the study of visual neurons. *Vision Research*, 37:2455–2464.

[Ringach et al., 2002] Ringach, D., Hawken, M., and Shapley, R. (2002). Receptive field structure of neurons in monkey primary visual cortex revealed by stimulation with natural image sequences. *Jouranl of Vision*, 2:12–24.

[Rinott, 1976] Rinott, Y. (1976). On convexity of measures. *Annals of Probability*, 4:1020–1026.

[Ritov and Bickel, 1990] Ritov, Y. and Bickel, P. (1990). Achieving information bounds in non- and semi-parametric models. *Annals of Statistics*, 18:925–938.

[Ruderman and Bialek, 1994] Ruderman, D. and Bialek, W. (1994). Statistics of natural images: scaling in the woods. *Physics Review Letters*, 73:814–817.

[Rust et al., 2003] Rust, N., Schwartz, O., Movshon, A., and Simoncelli, E. (2003). Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey v1 directionally selective neurons. *Submitted to CNS03 meeting*.

[Sahani, 1997] Sahani, M. (1997). Interactively exploring a neural code by active learning. Presented at NIC97 meeting, Snowbird, Utah; available at http://www.gatsby.ucl.ac.uk/ maneesh/conferences/nic97/poster/home.html.

[Schervish, 1995] Schervish, M. (1995). *Theory of statistics*. Springer-Verlag, New York.

[Schwartz et al., 2002] Schwartz, O., Chichilnisky, E., and Simoncelli, E. (2002). Characterizing neural gain control using spike-triggered covariance. *NIPS*, 14.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 7:461–464.

[Serfling, 1980] Serfling, R. (1980). *Approximation theorems of mathematical statistics*. Wiley, New York.

[Serruya et al., 2002] Serruya, M., Hatsopoulos, N., Paninski, L., Fellows, M., and Donoghue, J. (2002). Instant neural control of a movement signal. *Nature*, 416:141–142.

[Shao and Hahn, 1995] Shao, Y. and Hahn, M. (1995). Limit theorems for the logarithm of sample spacings. *Statistics and Probability Letters*, 24:121–32.

[Shapley and Victor, 1986] Shapley, R. and Victor, J. (1986). Hyperacuity in cat retinal ganglion cells. *Science*, 231:999–1002.

[Sharpee et al., 2003] Sharpee, T., Bialek, W., and Rust, N. (2003). Maximally informative dimensions: analyzing neural responses to natural signals. *Submitted to Neural Computation*.

[Shoham et al., 2003] Shoham, S., Fellows, M., Hatsopoulos, N., Paninski, L., Donoghue, J., and Normann, R. (2003). Optimal decoding for a primary motor cortical brain-computer interface. *Submitted*.

[Simmons, 1979] Simmons, J. (1979). Perception of echo phase in bat sonar. *Science*, 204:1336–1338.

[Simoncelli, 1999] Simoncelli, E. (1999). Modeling the joint statistics of images in the wavelet domain. In *Proc. SPIE, 44th Annual Meeting*, pages 188–195.

[Simoncelli and Heeger, 1998] Simoncelli, E. P. and Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761.

[Steele, 1986] Steele, J. (1986). An Efron-Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14:753–758.

[Strong et al., 1998] Strong, S. Koberle, R., de Ruyter van Steveninck R., and Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters*, 80:197–202.

[Talagrand, 1995] Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. IHES*, 81:73–205.

[Theunissen et al., 2001] Theunissen, F., David, S., Singh, N., Hsu, A., Vinje, W., and Gallant, J. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, 12:289–316.

[Tishby et al., 1999] Tishby, N., Pereira, F., and Bialek, W. (1999). The information bottleneck method. In *Proceedings 37th Allerton Conference on Communication, Control, and Computing*.

[Touryan et al., 2002] Touryan, J., Lau, B., and Dan, Y. (2002). Isolation of relevant visual features from random stimuli for cortical complex cells. *Journal of Neuroscience*, 22:10811–10818.

[Treves and Panzeri, 1995] Treves, A. and Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7:399–407.

[Tsodyks et al., 1999] Tsodyks, M., Kenet, T., Grinvald, A., and Arieli, A. (1999). Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286:1943–1946.

[Tzanakou et al., 1979] Tzanakou, E., Michalak, R., and Harth, E. (1979). The alopex process: Visual receptive fields by response feedback. *Biological Cybernetics*, 35:161–174.

[van der Vaart, 1998] van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge.

[van der Vaart and Wellner, 1996] van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes*. Springer-Verlag, New York.

[Victor, 2000a] Victor, J. (2000a). Asymptotic bias in information estimates and the exponential (bell) polynomials. *Neural Computation*, 12:2797–2804.

[Victor, 2000b] Victor, J. (2000b). How the brain uses time to represent and process visual information. *Brain Research*, 886:33–46.

[Victor, 2002] Victor, J. (2002). Binless strategies for estimation of information from neural data. *Physical Review E*, 66:51903–51918.

[Wandell, 1995] Wandell, B. (1995). *Foundations of Vision*. Sinauer, Boston.

[Watson and Fitzhugh, 1990] Watson, A. and Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception and Psychophysics*, 47:87–91.

[Watson and Pelli, 1983] Watson, A. and Pelli, D. (1983). Quest: a bayesian adaptive psychophysical method. *Perception and Psychophysics*, 33:113–120.

[Watson, 1980] Watson, G. (1980). *Approximation theory and numerical methods*. Wiley, Boston.

[Weiss et al., 2002] Weiss, Y., Simoncelli, E., and Adelson, E. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5:598–604.

[Wolpert and Wolf, 1995] Wolpert, D. and Wolf, D. (1995). Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52:6841–6854.

[Zhang et al., 1998] Zhang, K., Ginzburg, I., McNaughton, B., and Sejnowski, T. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *Journal of Neurophysiology*, 79:1017–1044.