Two problems from neural data analysis:

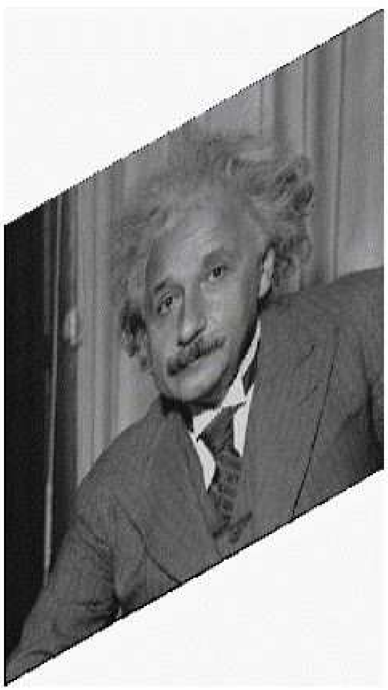
Sparse entropy estimation and efficient adaptive experimental design

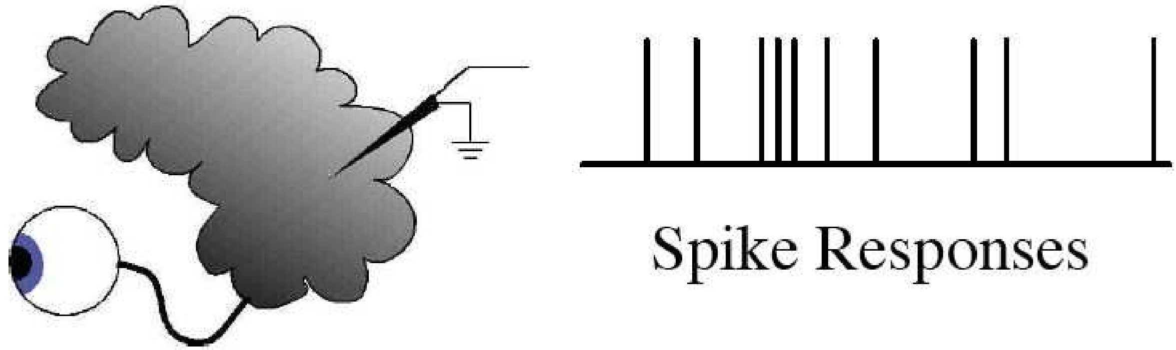
Liam Paninski

Department of Statistics and Center for Theoretical Neuroscience
Columbia University
http://www.stat.columbia.edu/∼liam
liam@stat.columbia.edu

May 9, 2006

The fundamental question in neuroscience





The **neural code**: what is $P(response \mid stimulus)$?

Main question: how to estimate $P(r|s)$ from (sparse) experimental data?

Curse of dimensionality

Both stimulus and response can be very high-dimensional.

Stimuli:

- images
- sounds
- time-varying behavior

Responses:

- observations from single or multiple simultaneously-recorded point processes

Avoiding the curse of insufficient data

1: Estimate some functional $f(p)$ instead of full joint $p(r, s)$

— information-theoretic functionals

2: Select stimuli more efficiently

— optimal experimental design

3: Improved nonparametric estimators

— minimax theory for discrete distributions under KL loss

(*4: Parametric approaches; connections to biophysical models*)

Part 1: Estimation of information

Many central questions in neuroscience are inherently information-theoretic:

- What inputs are most reliably encoded by a given neuron?
- Are sensory neurons optimized to transmit information about the world to the brain?
- Do noisy synapses limit the rate of information flow from neuron to neuron?

Quantification of “information” is fundamental problem.

(...interest in neuroscience but also physics, telecommunications, genomics, etc.)

Shannon mutual information

$$I(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} dp(x,y) \log \frac{dp(x,y)}{dp(x) \times p(y)}$$

Information-theoretic justifications:

- invariance
- "uncertainty" axioms
- data processing inequality
- channel and source coding theorems

But obvious open experimental question:

- is this computable for real data?

How to estimate information

I very hard to estimate in general...

... but lower bounds are easier.

Data processing inequality:

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

Suggests a sieves-like approach.

Discretization approach

Discretize $X, Y \to X_{disc}, Y_{disc}$, estimate

$$I_{discrete}(X;Y) = I(X_{disc};Y_{disc})$$

- Data processing inequality $\implies I_{discrete} \leq I$
- $I_{discrete} \nearrow I$ as partition is refined

Strategy: refine partition as samples N increases; if number of bins m doesn't grow too fast, $\hat{I} \to I_{discrete} \nearrow I$

Completely nonparametric, but obvious concerns:

- Want $N >> m(N)$ samples, to "fill in" histograms $p(x, y)$
- How large is bias, variance for fixed m?

Bias is major problem

$$\hat{I}_{MLE}(X;Y) = \sum_{x=1}^{m_x} \sum_{y=1}^{m_y} \hat{p}_{MLE}(x,y) \log \frac{\hat{p}_{MLE}(x,y)}{\hat{p}_{MLE}(x)\hat{p}_{MLE}(y)}$$

$$\hat{p}_{MLE}(x) = p_N(x) = \frac{n(x)}{N} \quad \text{(empirical measure)}$$

Fix $p(x,y), m_x, m_y$ and let sample size $N \to \infty$. Then:

- Bias($\hat{I}_{MLE}$) : $\sim -(m_x - m_y + m_x m_y - 1)/2N$.
- Variance($\hat{I}_{MLE}$) : $\sim (\log m)^2/N$; dominated by bias if $m = m_x m_y$ large.
- No unbiased estimator exists.

(Miller, 1955; Paninski, 2003)

Convergence of common information estimators

Result 1: If $N/m \to \infty$, $\hat{I}_{MLE}$ and related estimators universally almost surely consistent.

Converse: if $N/m \to c < \infty$, $\hat{I}_{MLE}$ and related estimators typically converge to *wrong* answer almost surely. (Asymptotic bias can often be computed explicitly.)

Implication: if N/m small, large bias although errorbars vanish, even if "bias-corrected" estimators are used (Paninski, 2003).

Estimating information on m bins with fewer than m samples

Result 2: A new estimator that is uniformly consistent as $N \to \infty$ even if $N/m \to 0$ (albeit sufficiently slowly)

Error bounds good for all underlying distributions: estimator works well even in *worst* case

Interpretation: information is strictly easier to estimate than p!
(Paninski, 2004)

Derivation of new estimator

Suffices to develop good estimator of discrete entropy:

$$I_{discrete}(X;Y) = H(X_{disc}) + H(Y_{disc}) - H(X_{disc}, Y_{disc})$$

$$H(X) = -\sum_{x=1}^{m_x} p(x) \log p(x)$$

Derivation of new estimator

Variational idea: choose estimator that minimizes upper bound on error over

$$\mathcal{H} = \{\hat{H} : \hat{H}(p_N) = \sum_i g(p_N(i))\} \quad (p_N = \text{empirical measure})$$

Approximation-theoretic (binomial) bias bound

$$\max_p Bias_p(\hat{H}) \leq B^*(\hat{H}) \equiv m \cdot \max_{0 \leq p \leq 1} \left| -p \log p - \sum_{j=0}^{N} g(\frac{j}{N}) B_{N,j}(p) \right|$$

McDiarmid-Steele bound on variance

$$\max_p Var_p(\hat{H}) \leq V^*(\hat{H}) \equiv N \max_j \left| g(\frac{j}{N}) - g(\frac{j-1}{N}) \right|^2$$

Derivation of new estimator

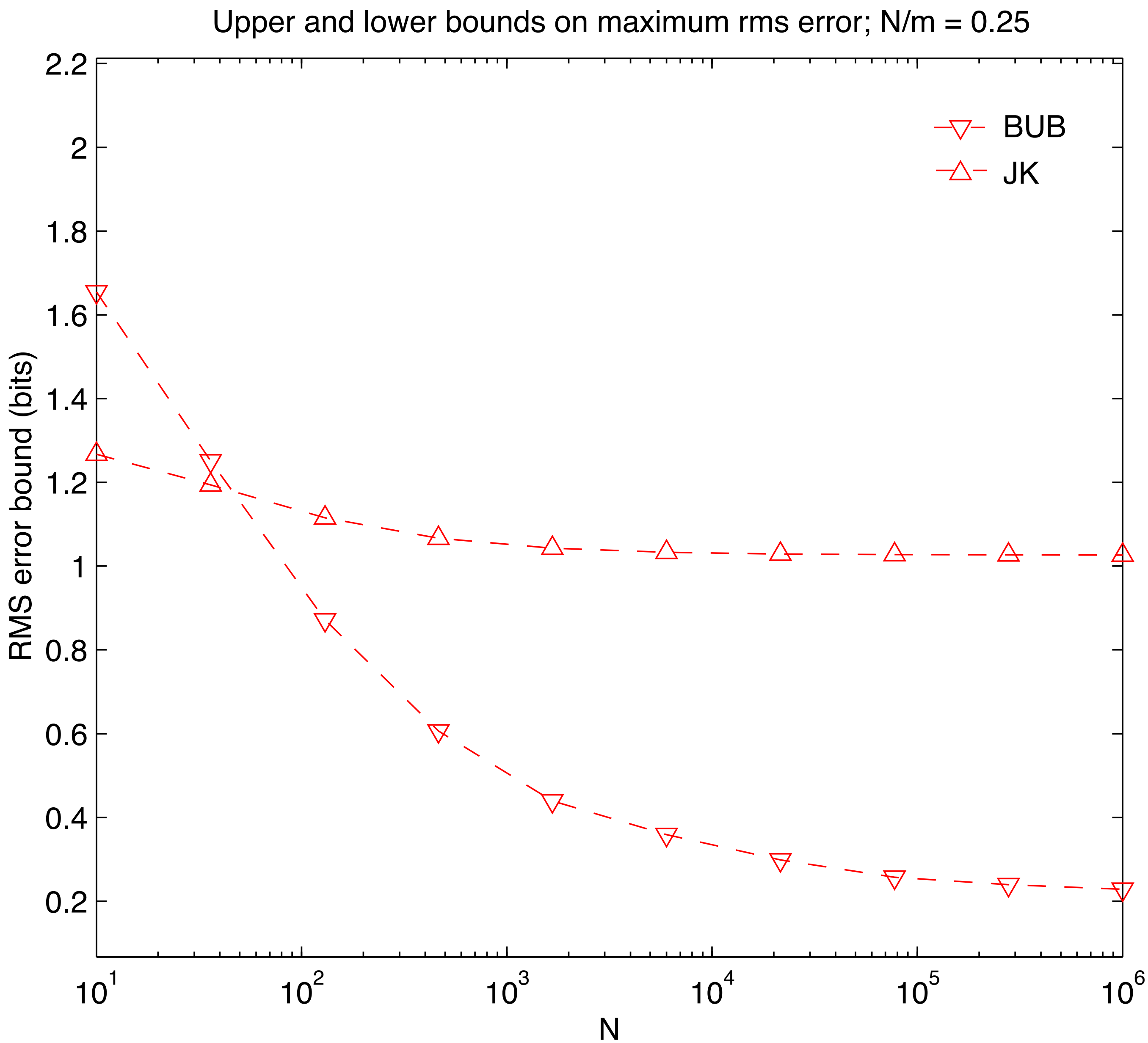
Choose estimator to minimize (convex) error bound over (convex) space $\mathcal{H}$:

$$\hat{H}_{BUB} = \text{argmin}_{\hat{H} \in \mathcal{H}} \, [B^*(\hat{H})^2 + V^*(\hat{H})].$$

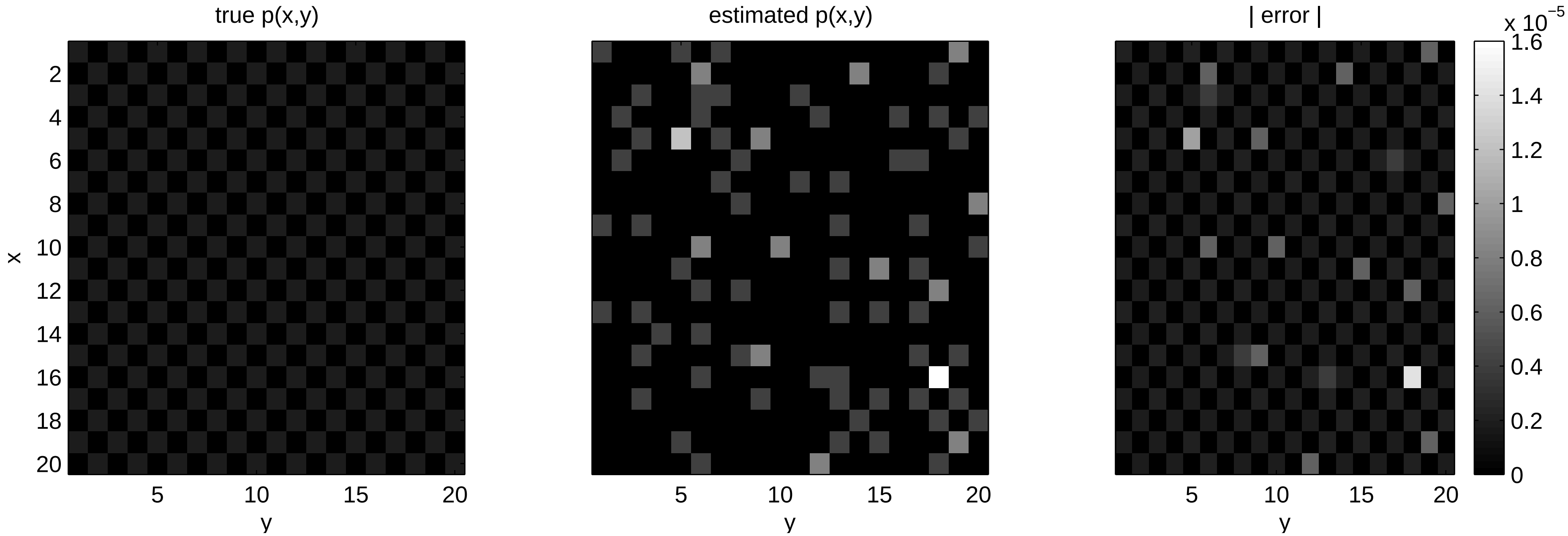
Optimization of convex functions on convex parameter spaces is computationally tractable by simple descent methods

Consistency proof involves Stone-Weierstrass theorem, penalized polynomial approximation theory in Poisson limit $N/m \to c$.

Error comparisons: upper and lower bounds



Undersampling example



$m_x = m_y = 1000; N/m_{xy} = 0.25$

$\hat{I}_{MLE} = 2.42$ bits

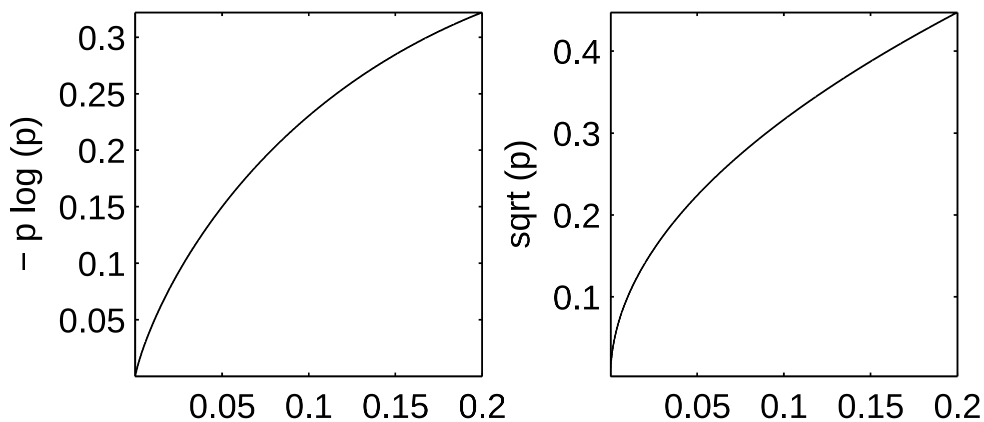
"bias-corrected" $\hat{I}_{MLE} = -0.47$ bits

$\hat{I}_{BUB} = \mathbf{0.74}$ bits; conservative (worst-case RMS upper bound) error: ± 0.2 bits

true $I(X;Y) = \mathbf{0.76}$ bits

Shannon ($-p \log p$) is special

Obvious conjecture: $\sum_i p_i^\alpha, 0 < \alpha < 1$ (Renyi entropy) should behave similarly.



Result 3: Surprisingly, not true: no estimator can uniformly estimate $\sum_i p_i^\alpha$, $\alpha \leq 1/2$, if $N \sim m$ (Paninski, 2004).

In fact, need $N > m^{(1-\alpha)/\alpha}$: smaller $\alpha \implies$ more data needed. (Proof via Bayesian lower bounds on minimax error.)

Directions

- KL-minimax estimation of full distribution in sparse limit $N/m \to 0$ (Paninski, 2005b)
- Continuous (unbinned) entropy estimators: similar result holds for kernel density estimates
- Sparse testing for uniformity: much easier than estimation ($N \gg m^{1/2}$ suffices)
- Open questions: $1/2 < \alpha < 1$? Other functionals?

Part 2: Adaptive optimal design of experiments

Assume:

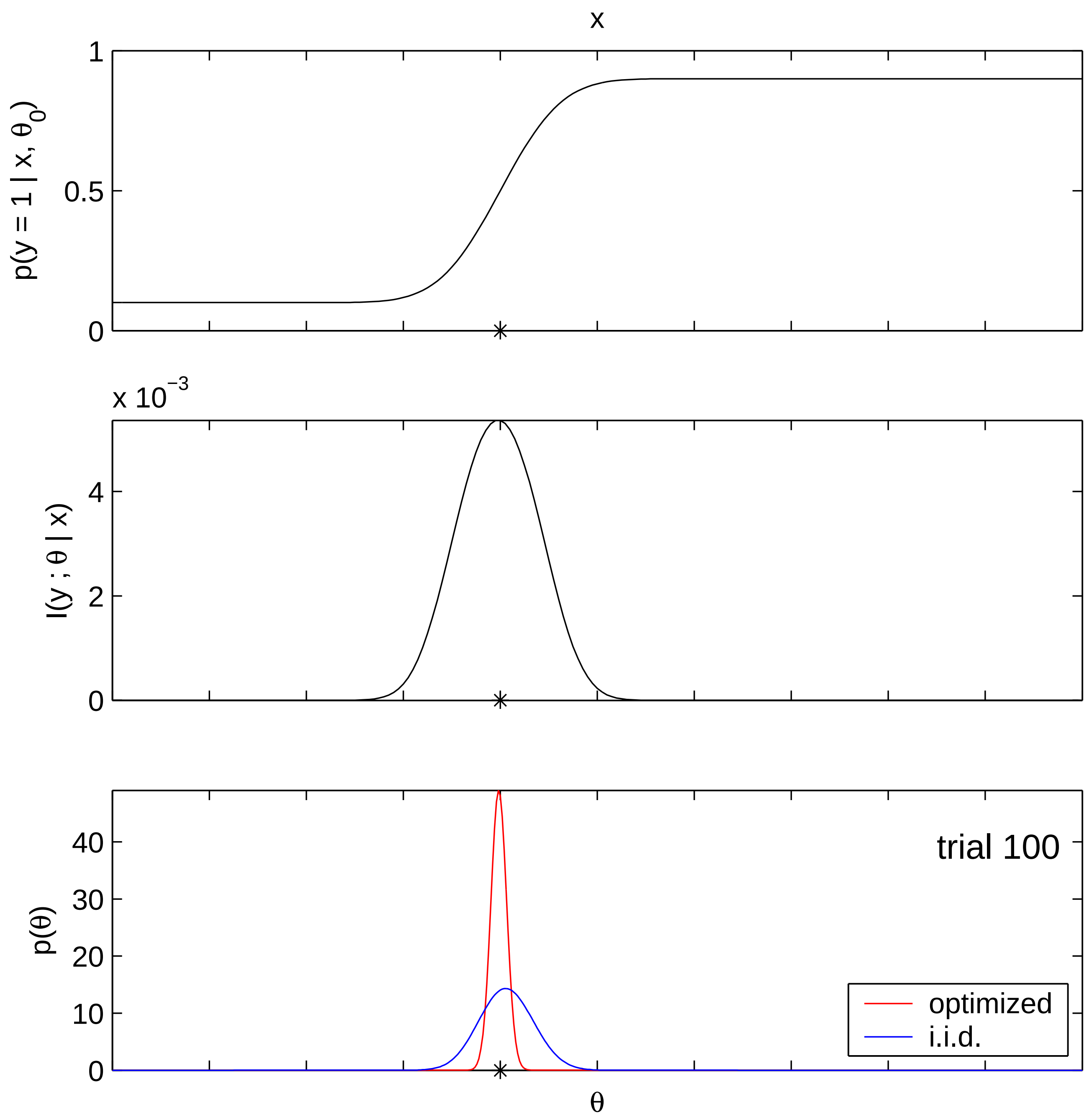
- parametric model $p_\theta(y|\vec{x})$ on outputs y given inputs $\vec{x}$
- prior distribution $p(\theta)$ on finite-dimensional model space

Goal: estimate θ from experimental data

Usual approach: draw stimuli i.i.d. from fixed $p(\vec{x})$

Adaptive approach: choose $p(\vec{x})$ on each trial to maximize $I(\theta; X)$ (e.g. "staircase" methods).

Snapshot: one-dimensional simulation



Main result

Under regularity conditions, a posterior CLT holds (Paninski, 2005a):

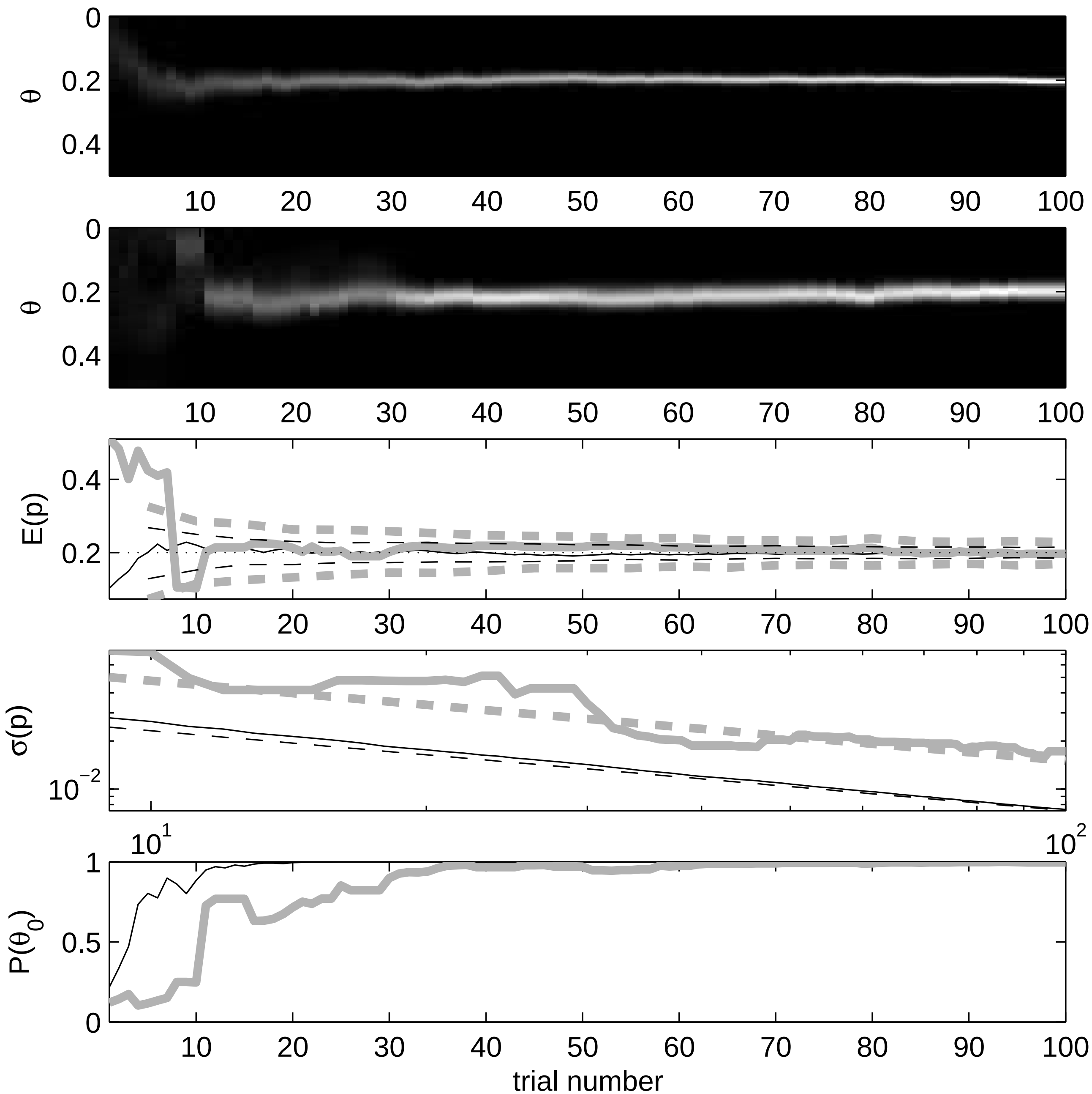
$$p_N\left(\sqrt{N}(\theta - \theta_0)\right) \to \mathcal{N}(\mu_N, \sigma^2); \quad \mu_N \sim \mathcal{N}(0, \sigma^2)$$

- $(\sigma^2_{iid})^{-1} = E_x(I_x(\theta_0))$
- $(\sigma^2_{info})^{-1} = \text{argmax}_{C \in co(I_x(\theta_0))} \log |C|$

$\implies \sigma^2_{iid} > \sigma^2_{info}$ unless $I_x(\theta_0)$ is constant in x

$co(I_x(\theta_0))$ = convex closure (over x) of Fisher information matrices $I_x(\theta_0)$. ($\log |C|$ strictly concave: maximum unique.)

Illustration of theorem



Technical details

Stronger regularity conditions than usual to prevent "obsessive" sampling and ensure consistency.

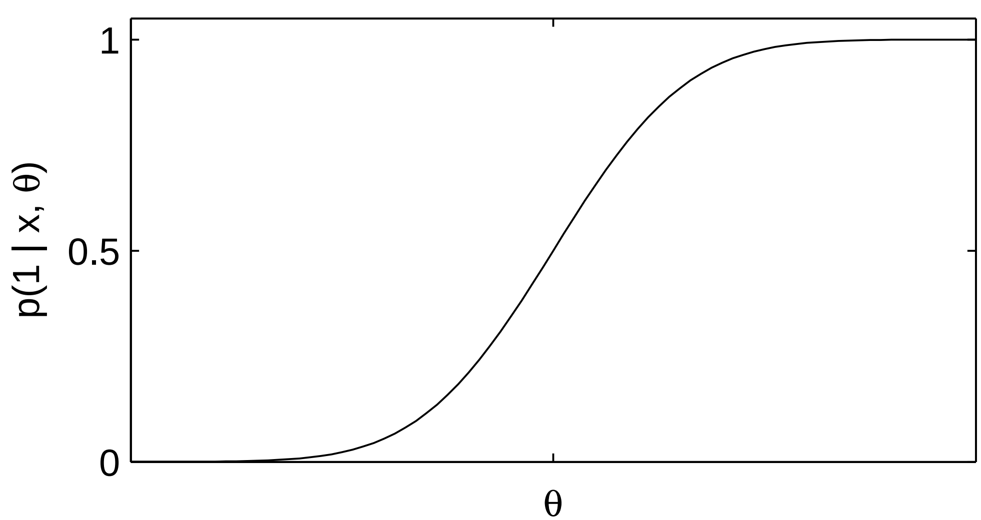
Significant complication: exponential decay of posteriors p_N off of neighborhoods of θ_0 does not necessarily hold.

Psychometric example

- stimuli x one-dimensional: intensity
- responses y binary: detect/no detect

$$p(1|x,\theta) = f((x-\theta)/a)$$

- scale parameter a (assumed known)
- want to learn threshold parameter θ as quickly as possible



Psychometric example: results

- variance-minimizing and info-theoretic methods asymptotically same

- just one unique function f^* for which $\sigma_{iid} = \sigma_{opt}$; for any other f, $\sigma_{iid} > \sigma_{opt}$

$$I_x(\theta) = \frac{(\dot{f}_{a,\theta})^2}{f_{a,\theta}(1 - f_{a,\theta})}$$

- f^* solves

$$\dot{f}_{a,\theta} = c\sqrt{f_{a,\theta}(1 - f_{a,\theta})}$$

$$f^*(t) = \frac{\sin(ct) + 1}{2}$$

- $\sigma^2_{iid}/\sigma^2_{opt} \sim 1/a$ for a small

Computing the optimal stimulus

Simple Poisson regression model for neural data:

$$y_i \sim Poiss(\lambda_i)$$

$$\lambda_i | \vec{x}_i, \vec{\theta} = f(\vec{\theta} \cdot \vec{x}_i)$$

Goal: learn $\vec{\theta}$ in as few trials as possible.

Problems:

- $\vec{\theta}$ is very high-dimensional; difficult to update $p(\vec{\theta}|\vec{x}_i, y_i)$, compute $I(\theta, y|\vec{x})$
- $\vec{x}$ is very high-dimensional; difficult to optimize $I(\theta, y|\vec{x})$

Efficient updating

Idea: Laplace approximation

$$p(\vec{\theta}|\{\vec{x}_i, y_i\}_{i\leq N}) \approx \mathcal{N}(\mu_N, C_N)$$

Justification:

- posterior CLT
- likelihood is log-concave, so posterior is also log-concave

$\Longrightarrow$ Updating μ_N, C_N is easy via Newton's method: $O(d^2)$ time

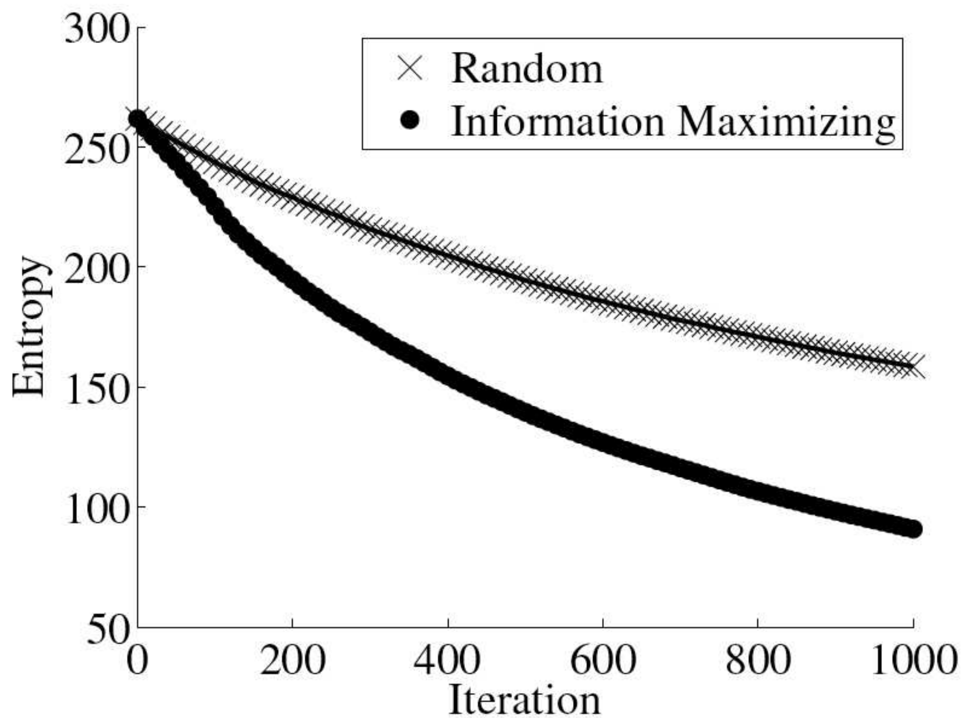
Efficient stimulus optimization

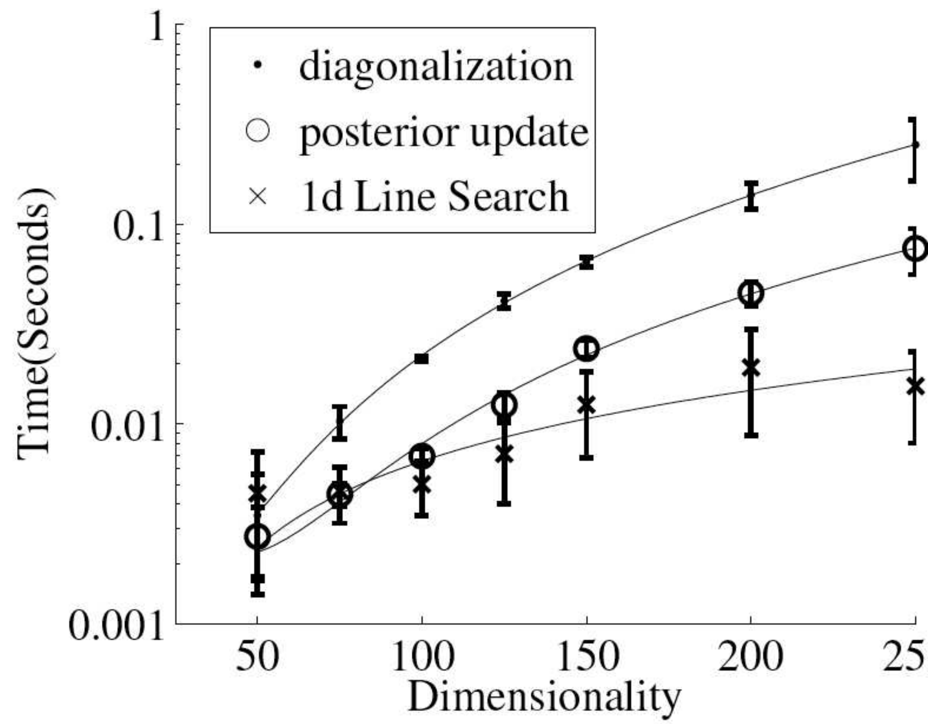
Sketch:

- Laplace approximation means Shannon information $\sim$ Fisher information
- Matrix perturbation theory simplifies nonlinear matrix problem
- Constraints on $||\vec{x}||_2$ reduce problem to eigenvalue problem followed by a numerical 1-dimensional optimization — much easier than full d-dimensional optimization!

$\Longrightarrow$ Computing optimal stimulus takes $O(d^3)$ time

Near real-time adaptive design





(Lewi et al., 2006)

References

Lewi, J., Butera, R., and Paninski, L. (2006). Efficient model-based information-theoretic design of experiments. *EMBS-06*.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press.

Miller, G. (1955). Note on the bias of information estimates. In *Information theory in psychology II-B*, pages 95–100.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253.

Paninski, L. (2004). Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50:2200–2203.

Paninski, L. (2005a). Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17:1480–1507.

Paninski, L. (2005b). Variational minimax estimation of discrete distributions under KL loss. *Advances in Neural Information Processing Systems*, 17.

Entropy bias bound

$$\begin{aligned}
Bias_p(\hat{H}) &= E_p(\hat{H}) - H(p) \\
&= \sum_{i=1}^{m} \left(p(i) \log p(i) + \sum_{j=0}^{N} g(\frac{j}{N}) B_{N,j}(p(i)) \right) \\
&\leq m \cdot \max_{0 \leq p \leq 1} | - p \log p - \sum_{j=0}^{N} g(\frac{j}{N}) B_{N,j}(p)|
\end{aligned}$$

- $B_{N,j}(p) = \binom{N}{j} p^j (1-p)^{N-j}$: polynomial in p
- If $\sum_j g(j) B_{N,j}(p)$ close to $-p \log p$ for all p, bias will be small

$\Longrightarrow$ standard uniform polynomial approximation theory

Back

Entropy variance bound

"Method of bounded differences" (McDiarmid, 1989): let $F(x_1, x_2, ..., x_N)$ be a function of N i.i.d. r.v.'s.

If any single x_i has small effect on F, i.e,

$$\sup |F(..., x, ...) - F(..., y, ...)| < c,$$

then

$$Var(F) < \frac{N}{4} c^2$$

(inequalities due to Azuma-Hoeffding, Efron-Stein, Steele, etc.).

Our case:

$$\hat{H} = \sum_i g(\frac{n(i)}{N})$$

$$\max_j \left| g(\frac{j}{N}) - g(\frac{j-1}{N}) \right| < c \implies Var\left(\sum_i g(\frac{n(i)}{N}) \right) \leq Nc^2$$

Back