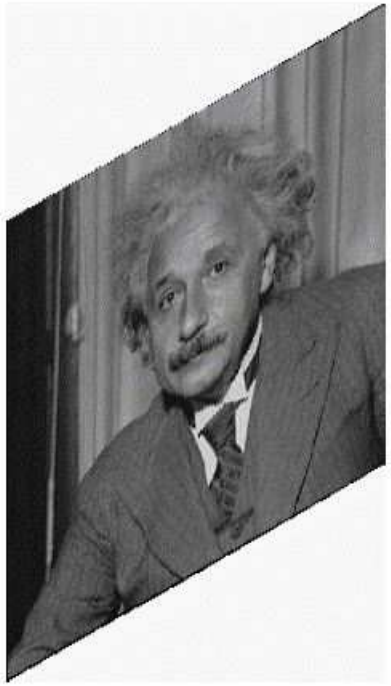
Nonparametric estimation of entropy and discrete distributions

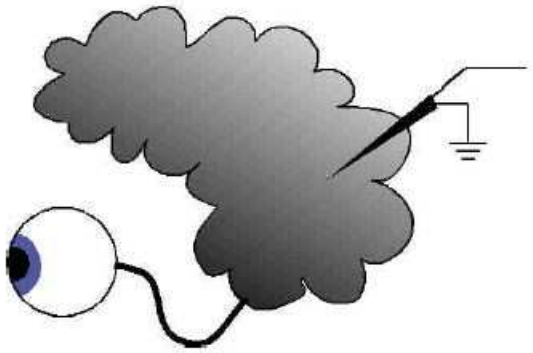
Liam Paninski

Department of Statistics
Columbia University
http://www.stat.columbia.edu/∼liam
liam@stat.columbia.edu

March 5, 2009

The fundamental question in neuroscience





Spike Responses

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 (neural activity)

Avoiding the curse of insufficient data

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

— information-theoretic functionals

2: Improved nonparametric estimators

— minimax theory for discrete distributions under KL loss

3: Select stimuli more efficiently

— optimal experimental design

(***4****: Parametric approaches*)

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

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

McDiarmid-Steele bound on variance

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

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

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:

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

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

Derivation of new estimator

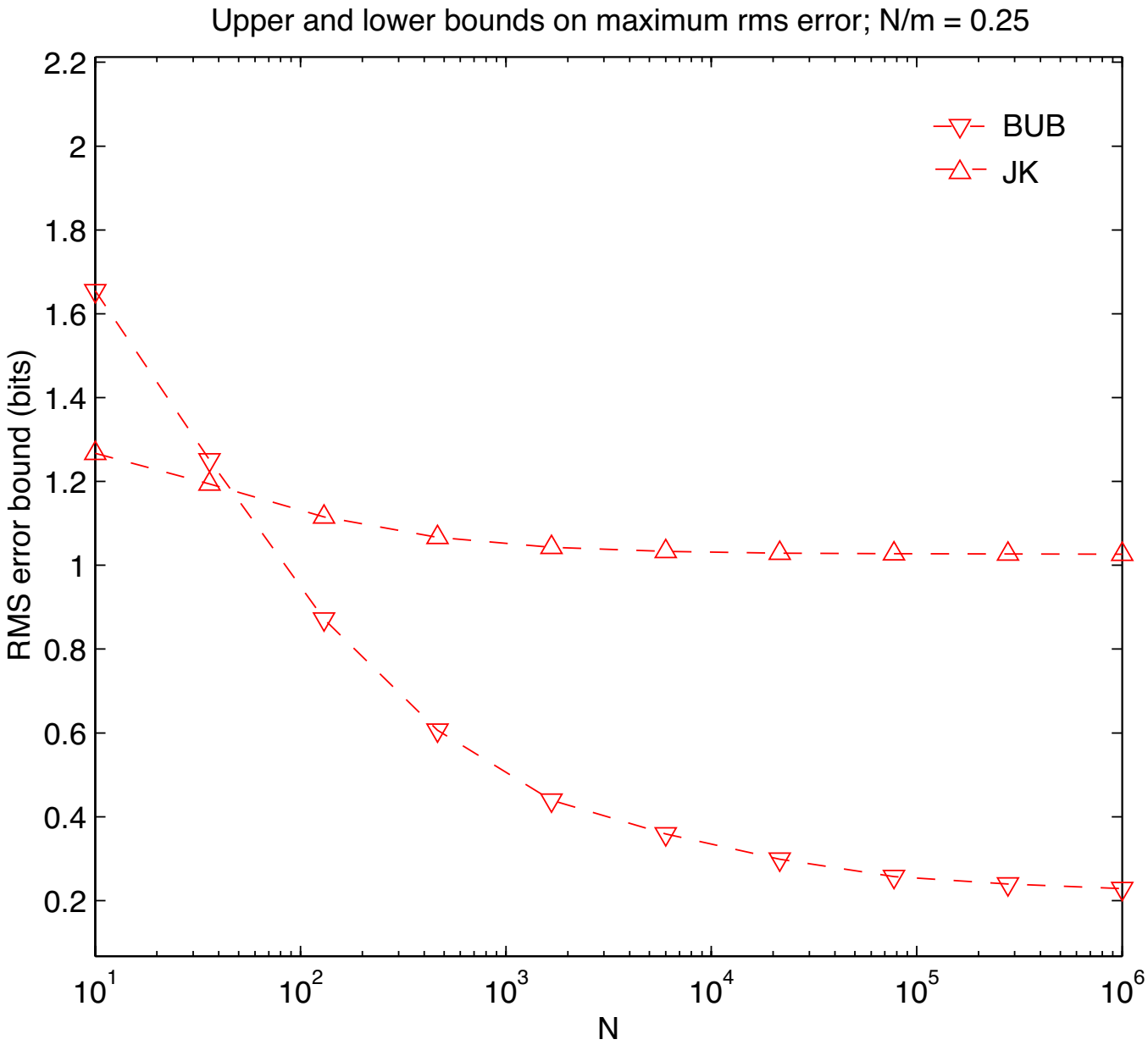
Choose estimator to minimize (convex) error bound over (convex) space $\mathcal{H}$:

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

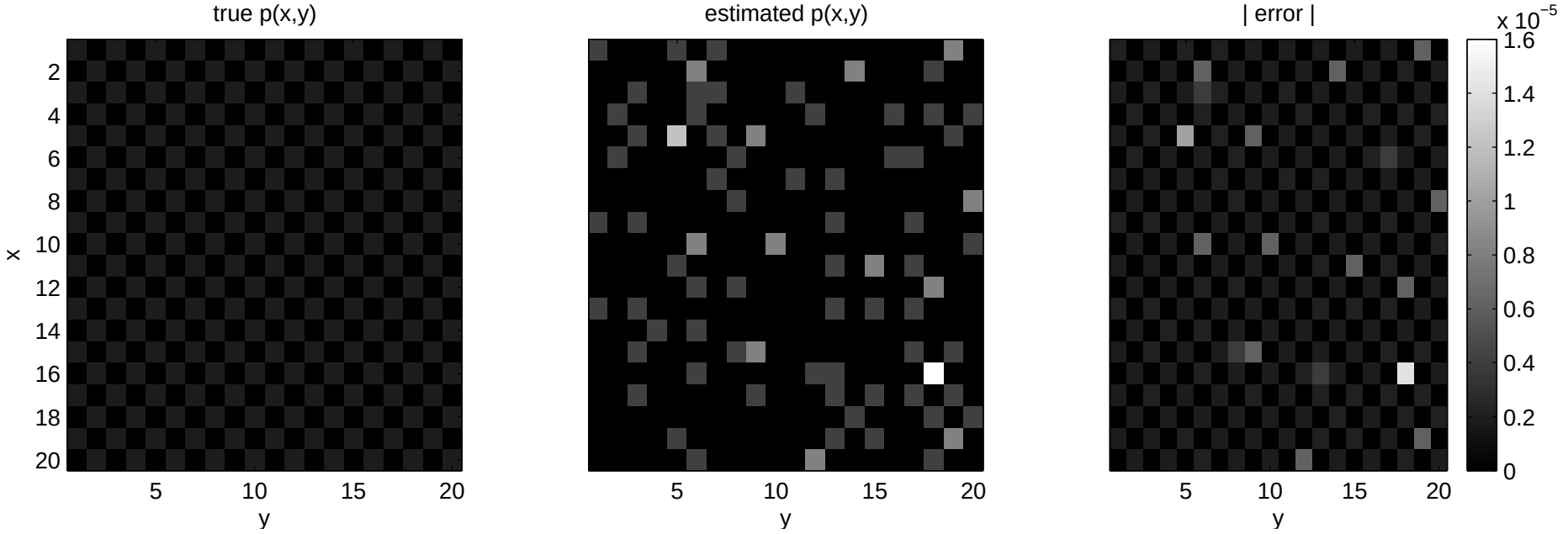
Optimization of convex functions on convex parameter spaces is computationally tractable by simple descent methods

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

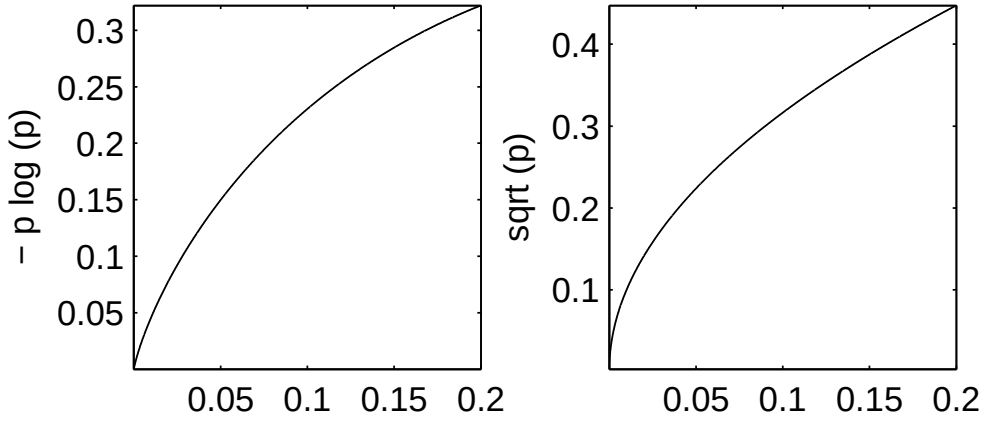
"bias-corrected" $\hat{I}_{MLE} = -0.47$ bits

$\hat{I}_{BUB} = \mathbf{0.74}$ 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.)

Sketch of lower bound

Idea: find two sequences of distributions p_0^N and p_1^N such that:

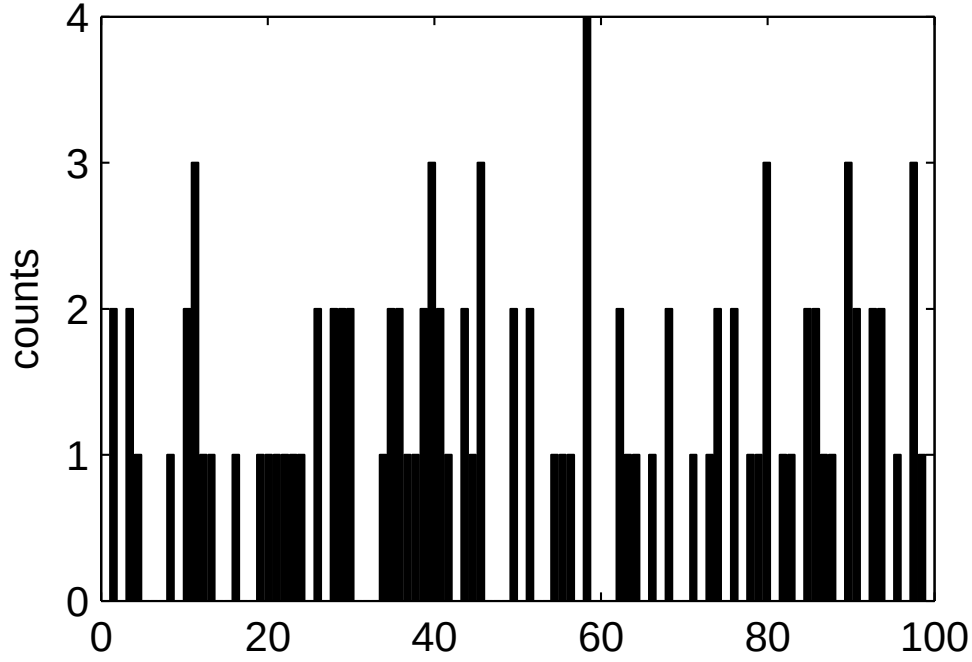
- $|F(p_0^N) - F(p_1^N)| > \epsilon > 0$
- p_0^N and p_1^N remain "statistically indistinguishable" (i.e., simple hypothesis test error remains bounded away from zero)

Here, p_0 places all of its mass on bin 1; p_1 places most of its mass on bin 1, but spreads some fraction uniformly on all other bins.

Directions

- $1/2 < \alpha < 1$? Other functionals?
- continuous (unbinned) entropy estimators: similar result holds for kernel density estimates
  (Paninski and Yajima, 2008)
- sparse hypothesis testing: much easier than estimation
  (Paninski, 2008)

Part 2: Estimating discrete distributions



- Want an estimator which works well even in worst case ("minimax" approach)
- Assume no knowledge of "topology" (smoothness); e.g. $p(word)$ in large dictionary

Of interest: sparse data case: N/m small (N = samples, m = number of bins).

Connections to entropy estimation

Information-theoretic error measure: Kullback-Leibler (relative entropy coding) loss $D_{KL}(p;\hat{p}) = \sum_i p_i \log(p_i/\hat{p}_i)$

What to do with unoccupied bins?

Sparse data case more interesting mathematically.

Methods turn out to be similar:

- Optimal approximation (Paninski, 2003)
- Dirichlet priors (Nemenman et al., 2002)

Upper bound idea

1. Derive upper bound on worst-case expected loss
2. Minimize upper bound over tractable class of estimators
3. Use estimator with best upper bound on worst-case loss

— want upper bound to be tight but tractable

Tractability:

1. Find bounds which are convex in the estimator
2. Allow estimators to range over a large convex space

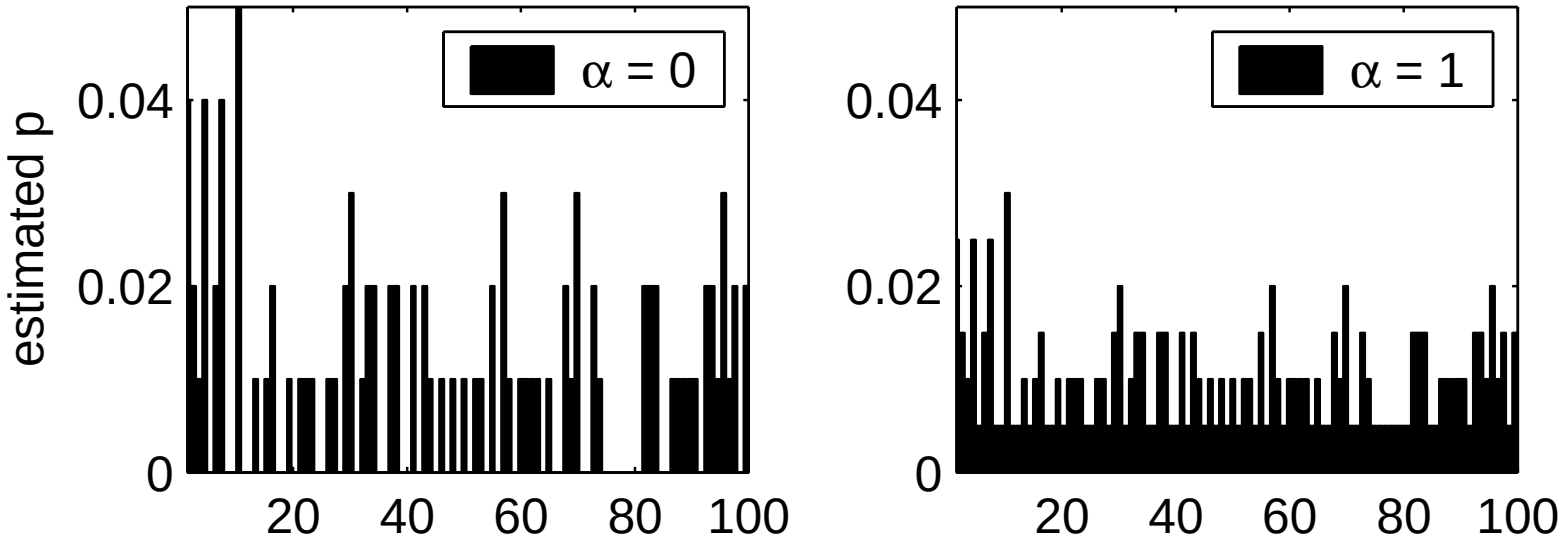
$\Longrightarrow$ no non-global local minima exist: descent methods work

Class of estimators

$$\hat{p}_i = \frac{g(n_i)}{\sum_{i=1}^{m} g(n_i)},$$

n_i number of samples observed in bin i

Example: "add-constant" estimators, $g_j \equiv g(j) = \frac{j+\alpha}{N+m\alpha}$, $\alpha > 0$



Basic upper bound

$$
\begin{aligned}
& E_{\vec{p}}\left(D_{KL}(\vec{p},\hat{p})\right) \\
&= E_{\vec{p}}\left(\sum_{i=1}^{m} p_i \log \frac{p_i}{\hat{p}_i}\right) \\
&= \sum_i \left(-H(p_i) + \sum_{j=0}^{N} (-\log g_j) p_i B_{N,j}(p_i)\right) + E_{\vec{p}}\left(\log \sum_{k=1}^{m} g(n_k)\right) \\
&\leq \sum_i \left(-H(p_i) + \sum_{j} (-\log g_j) p_i B_{N,j}(p_i)\right) + E_{\vec{p}}\left(-1 + \sum_{k} g(n_k)\right) \\
&= \sum_i f(p_i) = \sum_i -H(p_i) - p_i + \sum_j (g_j - p_i \log g_j) B_{N,j}(p_i).
\end{aligned}
$$

$$
\log p \leq p - 1; H(p) = -p \log p;\ B_{N,j}(p) = \binom{N}{j} p^j (1-p)^{N-j}
$$

Equality iff $\sum_k g(n_k)$ constant (e.g., add-constant estimator).

Two upper bounds

$$\sum_{i=1}^{m} f(p_i) \leq m \max_{0 \leq p \leq 1} f(p) :$$

tight in heavily-sampled limit.

$$\sum_{i} f(p_i) \leq \left(m \max_{0 \leq p \leq 1/m} f(p) \right) + \left(\max_{1/m \leq p \leq 1} \frac{f(p)}{p} \right) :$$

tight in sparse-data limit.

Minimizing bounds = polynomial approximation problem, as in entropy estimation case (Paninski, 2003; Paninski, 2004).

Note: bounds are convex in $g_j \implies$ easy to minimize!

Lower bounds

Compare upper bounds to some well-defined optimum: lower bound on worst-case error of *any* estimator.

Derive family of bounds indexed by some parameter α, then optimize over α.

Key idea: average (Bayesian) loss $\leq$ maximum (worst-case) loss.

Dirichlet lower bounds

Dirichlet priors are convenient (Cover, 1972; Krichevsky, 1998; Nemenman et al., 2002):

$$P(p) = \frac{1}{Z(\vec{\alpha})} \prod_i p_i^{\alpha_i - 1}$$

1. Compute the average error under any Dirichlet distribution; can be done analytically for any $N, m, \vec{\alpha}$.
2. Maximize over the possible Dirichlet parameters α (i.e., find "least favorable" Dirichlet prior) to obtain tightest bound
3. Simplification: restrict $\vec{\alpha}$ to be constant, $\vec{\alpha} = (\alpha, \alpha, \ldots, \alpha)$ (1-D maximization)
4. Bound achieved by add-α estimator.

Asymptotic analysis

Proposition 1. *Any add-α estimator, $\alpha > 0$, is uniformly KL-consistent if $N/m \to \infty$.*

Note: N/m is allowed to tend to infinity arbitarily slowly.

Proposition 2 (Converse)**.** *No estimator is uniformly KL-consistent if* $\limsup N/m < \infty$.

— Contrast with entropy estimation, where consistent estimators do exist in this regime (conjectured by (Nemenman et al., 2002; Paninski, 2003); proven in (Paninski, 2004)).

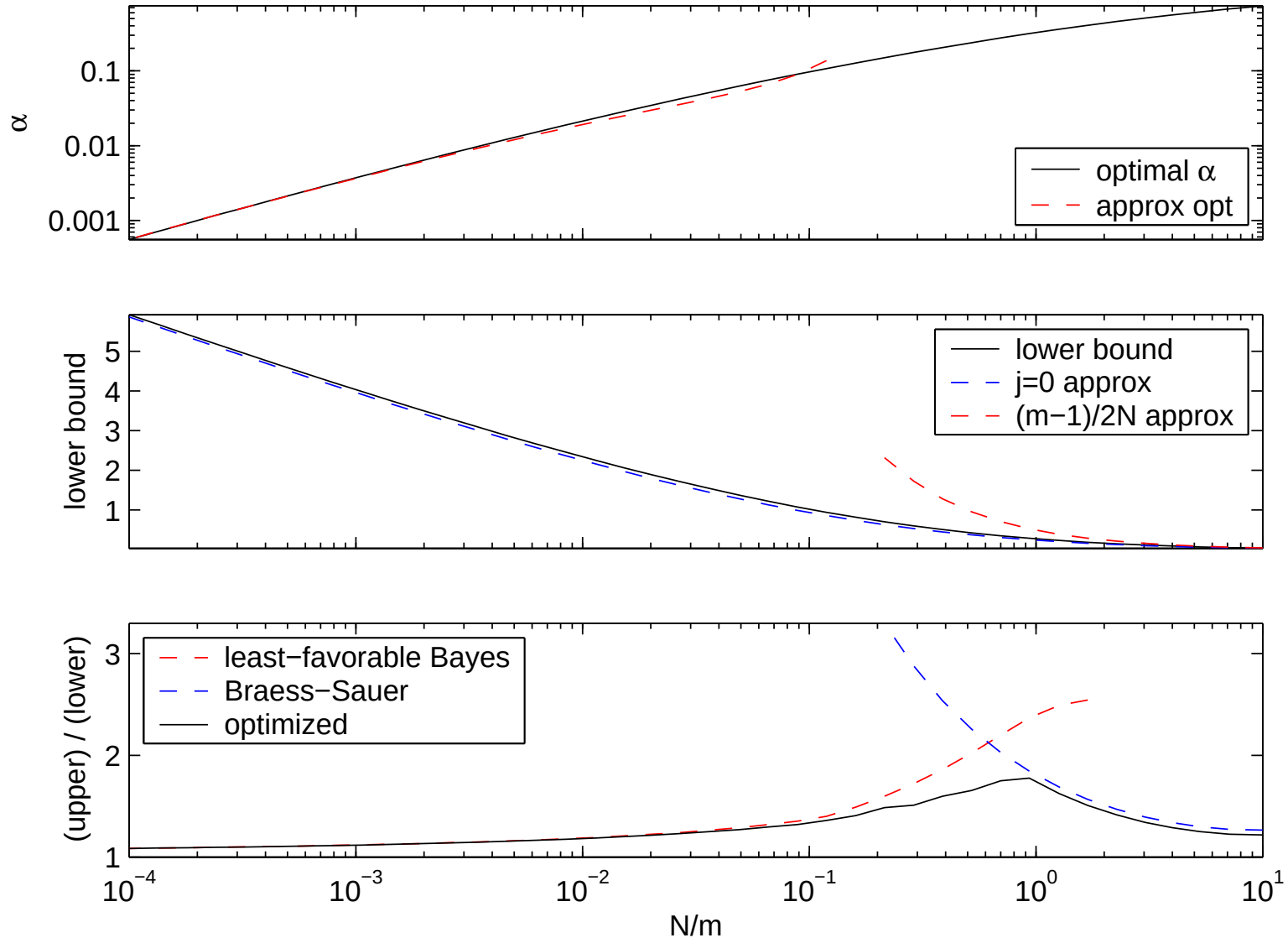
$\implies$ entropy of p is strictly easier to estimate than p!

Main result

In data-sparse regime $c \equiv N/m \to 0$, add-α estimator with $\alpha = -c \log c$ is optimal.

Proposition 3. *The least-favorable Dirichlet parameter is given by $H(c)$ as $c \to 0$; the corresponding add-$H(c)$ estimator also asymptotically minimizes the upper bound. The maximal and average error behave as $-\log(c)(1 + o(1))$ for $c \to 0$.*

Illustration of bounds



— asymptotically tight as $c \to 0$, $c \to \infty$

— always tight within a factor of 2

Summary

- New upper and lower bounds on discrete estimation error
- Useful applications of (convex) variational idea
- Proved asymptotic tightness of bounds
- Numerically, bounds turn out to be fairly tight
- Optimal sparse estimator: add-$|c \log c|$
- See (Paninski, 2005) for details.

References

Cover, T. (1972). Admissibility properties of Gilbert's encoding for unknown source probabilities. *IEEE Transactions on Information Theory*, 18:216–217.

Krichevsky, R. (1998). Laplace's law of succession and universal encoding. *IEEE Transactions on Information Theory*, 44:296–303.

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.

Nemenman, I., Shafee, F., and Bialek, W. (2002). Entropy and inference, revisited. *NIPS*, 14.

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

Paninski, L. (2008). A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54:4750–4755.

Paninski, L. and Yajima, M. (2008). Undersmoothed kernel entropy estimators. *IEEE Transactions on Information Theory*, 54:4384–4388.