Maximum Entropy Sampling

Jon Lee

Industrial and Operations Engineering Department
University of Michigan
Ann Arbor, Michigan

20 March 2012
Examples

- Coverage functions
  - Underlying finite universe $U$
  - Each $i \in N$ corresponds to a set $T_i \subset U$
  - $f(S) = |\bigcup_{i \in S} T_i|$, for $S \subset N$
  - Properties: monotone and nonnegative
Examples

- **Coverage functions**
  - Underlying finite universe $U$
  - Each $i \in N$ corresponds to a set $T_i \subset U$
  - $f(S) = |\bigcup_{i \in S} T_i|$, for $S \subset N$
  - Properties: monotone and nonnegative

- **Cut functions**
  - (Di)Graph $G = (N, E)$ on vertex set $N$
  - $f(S) = |\delta_{(+)}(S)|$, for $S \subset N$
  - Property: cut functions are nonnegative
  - Property: cut function in undirected graphs are symmetric
Examples

- Coverage functions
  - Underlying finite universe $U$
  - Each $i \in N$ corresponds to a set $T_i \subset U$
  - $f(S) = |\bigcup_{i \in S} T_i|$, for $S \subset N$
  - Properties: monotone and nonnegative

- Cut functions
  - (Di)Graph $G = (N, E)$ on vertex set $N$
  - $f(S) = |\delta_{(+)}(S)|$, for $S \subset N$
  - Property: cut functions are nonnegative
  - Property: cut function in undirected graphs are symmetric

- Matroids / matroid intersection (e.g., trees / assignments)
  - $f(S) := r(S)$
  - $f(S) := r_1(S) + r_2(N \setminus S)$
    (min of $f =$ maximum size of a set “independent” in $M_1$ and $M_2$)
  - Properties: monotone and nonnegative
Examples

- **Coverage functions**
  - Underlying finite universe $U$
  - Each $i \in N$ corresponds to a set $T_i \subset U$
  - $f(S) = |\bigcup_{i \in S} T_i|$, for $S \subset N$
  - Properties: monotone and nonnegative

- **Cut functions**
  - (Di)Graph $G = (N, E)$ on vertex set $N$
  - $f(S) = |\delta_+(S)|$, for $S \subset N$
  - Property: cut functions are nonnegative
  - Property: cut function in undirected graphs are symmetric

- **Matroids / matroid intersection** (e.g., trees / assignments)
  - $f(S) := r(S)$
  - $f(S) := r_1(S) + r_2(N \setminus S)$
    (min of $f =$ maximum size of a set “independent” in $M_1$ and $M_2$)
  - Properties: monotone and nonnegative

- **(Differential/continuous) entropy of a finite set of Gaussian random variables**
  - $f(S) := \log \det C[S]$, where $C[N]$ is positive semidefinite
  - Does not trivially have additional, nice properties
Information = Disorder

“Chance and chance alone has a message for us. Everything that occurs out of necessity, everything expected, repeated day in and day out, is mute. Only chance can speak to us. We read its message much as gypsies read the images made by coffee grounds at the bottom of a cup.”

- Milan Kundera (The Unbearable Lightness of Being)
“I propose to name the magnitude $S$ the entropy of the body from the Greek word ἑντροπία, a transformation. I have intentionally formed the word entropy so as to be as similar as possible to the word energy, since both these quantities, which are to be known by these names, as so nearly related to each other in their physical significance that a certain similarity in their names seemed to me advantageous ...”

— R. Clausius (1865)
Historical Highlights

- R. Clausius (1865) — “entropy” (also Carnot and Kelvin in their versions of the 2nd law of thermodynamics), arrow of time (“What then is time? If no one asks me, I know what it is. If I wish to explain it to him who asks, I do not know.” — St. Augustine)
- L. Boltzmann (1877) — statistical mechanics
- C. Shannon (1948) — information theory
- D. Blackwell (1951) — statistics
Entropy more recently...
and more...
Maximum-Entropy Sampling

\( N = \{1, 2, \ldots, n\} \)

Random \( Y_N = \{Y_j : j \in N\} \) with continuous density \( g_N \)

**Goal:** Choose \( S \subset N \), with \( |S| = s \), so that observing \( Y_S \) maximizes the “information” obtained about \( Y_N \).

**Entropy:** \( h(S) := -E[\ln g_S(Y_S)] \).
Motivation: Environmental Monitoring

- Sites of emission $\implies$ Causes
- Sites of deposition $\implies$ Effects*

* Clean Air Act of 1990 and its revisions mandate effects monitoring

National Acidic Deposition Program/
National Trends Network
nadp.sws.uiuc.edu

Precipitation collected weekly; analyzed for: Hydrogen (acidity as pH — ‘acid rain’), Sulfate, Nitrate, Ammonia, Chloride, Calcium, Magnesium, Potassium, Sodium
Wet vs. Dry
TPC 3000 (Yankee Environ. Sys.)
YES has US federal funding of $300K to develop a new prototype over 2 years

$3.5M federal funding for NTN ('99)

~ $150M total US federal funding for environmental monitoring ('99)

- much other monitoring focused on CO, NO$_2$, SO$_2$ and small particulate matter
Sulfate ion concentration, 1994

Sites not pictured:
AK01  0.2 mg/L
AK03  0.2 mg/L
PR20  0.8 mg/L

National Atmospheric Deposition Program/National Trends Network
http://nadp.sws.uiuc.edu
Sulfate ion concentration, 2002

Sites not pictured:
- AK01 0.2 mg/L
- AK03 0.3 mg/L
- CA95 2.3 mg/L
- HI99 1.1 mg/L
- VI01 0.9 mg/L

National Atmospheric Deposition Program/National Trends Network
http://nadp.sws.uiuc.edu
Sulfate ion concentration, 2008

Sites not pictured:
AK01 0.2 mg/L
AK03 0.1 mg/L
PR20 0.7 mg/L
VI01 0.6 mg/L

National Atmospheric Deposition Program/National Trends Network
http://nadp.sws.uiuc.edu
Nice Properties of Entropy

- **Submodularity ✓**: \( h(S \cup T) + h(S \cap T) \leq h(S) + h(T) \)
- The Gaussian distribution maximizes the entropy for a given covariance matrix \( C \)
- Gaussian case: \( h(S) = k_s + k \ln \det C[S, S] \)
- **Conditional Additivity**: 
  \[
  h(N) = \underbrace{\max}_{\text{max}} h(S) + \underbrace{\min}_{\text{min}} h(N \setminus S|S) 
  \]
  (justifies our objective function)
- **Change coordinate systems**: Entropy difference is \( \log \det(\text{Jacobian of transformation}) \)
- **Complementation**: 
  \[
  \ln \det C[S, S] = \ln \det C + \ln \det C^{-1}[N \setminus S, N \setminus S] 
  \]
Proposition [Ko, Lee, Queyranne]. The maximum-entropy sampling problem is **NP-Hard** (even for the Gaussian diagonally-dominant case)

**Proof:**

- **INDEPENDENT SET:** Does a simple undirected graph $G$ on $n$ vertices have an independent set of vertices of cardinality $s$?
- Let $C := A(G) + 3nI$

$$
\begin{pmatrix}
12 & 1 & 0 & 0 \\
1 & 12 & 1 & 1 \\
0 & 1 & 12 & 0 \\
0 & 1 & 0 & 12
\end{pmatrix}
$$
(KLQ) Branch . . .

- **Fixing $j$ out of $S$:**
  - Strike out row and column $j$: $C[N, N] \rightarrow C[N - j, N - j]$

- **Fixing $j$ in $S$:**
  - Schur complement of $C[j, j]$: $C[N, N] \rightarrow$
    

  (and solution/bounds are shifted by $\ln C[j, j]$.)
... and Bound

- **Lower bounds**: Greedy, local-search, rounding heuristics
- **Upper bounds**:
  - **Spectral based bounds**
    - Ko, Lee, Queyranne ’95 (original B&B and spectral bound)
    - Lee ’98 (extension to side constraints)
    - Hoffman, Lee & Williams ’01 (spectral partition bounds)
    - Lee, Williams ’03 (tightening HLW via ILP and matching)
    - Anstreicher, Lee ’04 (generalization of HLW)
    - Burer, Lee ’07 (another approach to computing the AL bound)
  - **NLP relaxation**
    - Anstreicher, Fampa, Lee & Williams ’96 (continuous NLP relaxation and parallel B&B)
Complementary Bounds (Anstreicher, Fampa, Lee, Williams)

\[ \ln \det C[S, S] = \ln \det C + \ln \det C^{-1}[N \setminus S, N \setminus S] \]

- So a maximum entropy $s$-subset of $N$ with respect to $C$ is the complement of a maximum entropy $(n - s)$-subset of $N$ with respect to $C^{-1}$
- So a bound on the complementary problem plus the entropy of the entire system is a bound on the original problem
- These complementary bounds can be quite effective
NLP Bound (AFLW)

\[
\max f(x) := \ln \det \left( \operatorname{Diag}(x_j^{p_j}) C \operatorname{Diag}(x_j^{p_j}) + \operatorname{Diag}(d_j x_j - d_j x_j^{p_j}) \right)
\]

subject to \( \sum_{j \in N} a_{ij} x_j \leq b_i, \forall i; \quad \iff \text{CONSTRAINTS} \)

\[
\sum_{j \in N} x_j = s;
\]

\( 0 \leq x_j \leq 1, \forall j, \)

where the constants \( d_j > 0 \) and \( p_j \geq 1 \) satisfy

\( d_j \leq \exp(p_j - \sqrt{p_j}), \) and \( \operatorname{Diag}(d_j) - C[N, N] \succeq 0. \)
NLP Bound, cont’d

\[
\max f(x) := \ln \det \left( \text{Diag}(x_j^{p_j}) \ C \ \text{Diag}(x_j^{p_j}) + \text{Diag}(d_j^{x_j} - d_j x_j^{p_j}) \right)
\]

For \( (1, 1, \ldots, 1, 0, 0, \ldots, 0) \)

- \( \text{Diag}(d_j^{x_j} - d_j x_j^{p_j}) = \text{Diag}(0, 0, \ldots, 0, 1, 1, \ldots, 1) \).
- \( \text{Diag}(x_j^{p_j}) \ C \ \text{Diag}(x_j^{p_j}) = \begin{pmatrix}
C[S, S] & 0 \\
0 & 0
\end{pmatrix} \)
NLP Bound: Properties

- **Concavity:** Assume $D \succeq C$, $p_j \geq 1$, $0 < d_j \leq \exp(p_j - \sqrt{p_j})$. Then $f$ is concave for $0 < x \leq e$

- **Dominance:** Assume that $p$ and $d$ satisfy the above, and $p' \geq p$. Let $f'$ be defined as above, but using $p'$ for $p$. Then $f'(x) \geq f(x) \forall 0 < x \leq e$

- **Scaling $C$ by $\gamma$ adds $s \ln(\gamma)$ to the obj.** Let $f_\gamma(x) := \ln \det \left( \gamma X^{p/2} (C - D) X^{p/2} + (\gamma D)^x \right) - s \ln(\gamma)$
  - **Scaling:** Assume $I \succeq D \succeq C$, $p = e$. Then $f_\gamma(x) \geq f(x) \forall 0 \leq x \leq e$, $e^T x = s$ and $0 < \gamma \leq 1$
  - Assume $D \succeq C$, $D \succeq I$. Then $f_\gamma(x) \geq f(x) \forall 0 < x \leq e$, $e^T x = s$ and $\gamma \geq 1$, where $p$ is chosen as above

These results give us some guidance for choosing the $p_j$, $d_j$ and $\gamma$
Spectral Bound (KLQ)

\[ z \leq \sum_{l=1}^{s} \ln \lambda_l(C) \]

- Determinant = product of eigenvalues.
- Eigenvalue interlacing.

\[
\begin{pmatrix}
\lambda_1 & \geq & \lambda'_1 \\
\lambda_2 & \geq & \lambda'_2 \\
\lambda_3 & \geq & \lambda'_3 \\
\vdots & & \vdots \\
\lambda_s & \geq & \lambda'_s
\end{pmatrix}
\]
Lagrangian Spectral Bound (Lee)

For handling linear side constraints

$$\min_{\pi \in \mathbb{R}^m_+} v(\pi)$$

where

$$v(\pi) := \left\{ \sum_{l=1}^{s} \ln \lambda_l (D^\pi C D^\pi) + \sum_{i \in M} \pi_i b_i \right\},$$

and $D^\pi$ is the diagonal matrix having

$$D^\pi_{jj} := \exp \left\{ -\frac{1}{2} \sum_{i \in M} \pi_i a_{ij} \right\}$$
Optimizing the Lagrangian Spectral Bound

- $v_\pi$ is convex (in $\pi$)
- $v_\pi$ is analytic when $\lambda_s (D^\pi C D^\pi) > \lambda_{s+1} (D^\pi C D^\pi)$
Let \( x^l \) be the eigenvector (of unit Euclidean norm) associated with \( \lambda_l \).

Define the continuous solution \( \tilde{x} \in \mathbb{R}^N \) by 
\[
\tilde{x}_j := \sum_{i=1}^{s} \left( x^l_j \right)^2, \quad \text{for } j \in N.
\]

Define \( \gamma \in \mathbb{R}^M \) by 
\[
\gamma_i := b_i - \sum_{j \in N} a_{ij} \tilde{x}_j.
\]

If \( \lambda_s > \lambda_{s+1} \), then \( \gamma \) is the gradient of \( f \) at \( \pi \).

Can incorporate this in a Quasi-Newton (or, with an expression for the Hessian, a Newton) method for finding the minimum. (Implemented using LBFGS-B (Zhu, Byrd, Nocedal) and a coarse line search)
Spectral Partition Bound (Hoffman, Lee, Williams)

Let $\mathcal{N} = \{N_1, N_2, ..., N_n\}$ denote a partition of $N$. Let $C'' = 0$ except for $C'[N_k, N_k] = C[N_k, N_k]$.

$$z \leq \sum_{l=1}^{s} \ln \lambda_l(C')$$

- Based on “Fischer’s Inequality”
- For $\mathcal{N} = \{\{1\}, \{2\}, \ldots, \{n\}\}$ we have “the diagonal bound”
- For $\mathcal{N} = \{N, \emptyset, \emptyset, \ldots, \emptyset\}$ we have the ordinary spectral bound
- As we partition $N$, the optimal value with respect to $C'$ cannot decrease, but the bound can decrease
ILP Bound (Lee, Williams)

Observation: Why calculate eigenvalue based bounds for small blocks of a partition? Just solve the small blocks exactly.

\[ x_k(i) = 1 \iff \text{pick } k \text{ elements from block } N_i \]

\[ g_s(N) := \max_{i=1}^{p} \sum_{k=1}^{\lfloor N_i \rfloor} \sum_{k=1}^{\lfloor N_i \rfloor} f_k(N_i) x_k(i) \]

s.t. \[ \sum_{k=1}^{\lfloor N_i \rfloor} x_k(i) \leq 1, \text{ for } i = 1, 2, \ldots, p; \]

\[ \sum_{i=1}^{p} \sum_{k=1}^{\lfloor N_i \rfloor} k x_k(i) = s \]

\[ x_k(i) \in \{0, 1\}, \text{ for } i = 1, 2, \ldots, p, \quad k = 1, 2, \ldots, \lfloor N_i \rfloor. \]
**ILP Bound, cont’d**

- Refines the spectral partition bound.
- Calculate via dynamic programming (assuming $|N_i|$ is bounded):
  - **Boundary conditions:**
    - $v_t(j) := -\infty$ when $\sum_{i=1}^j |N_i| < t \leq s$;
    - $v_0(0) := 0$.
    
    $$v_t(j) = \max_{0 \leq k \leq \min\{|N_j|, t\}} \left\{ f_k(N_j) + v_{t-k}(j-1) \right\}.$$ 

  Then $v_s(p) = g_s(N)$

- Can even calculate via Edmonds’ min-weight matching algorithm when $|N_i| \leq 2$. 
A **mask** is a (symmetric) $X \succeq 0$ having $\text{diag}(X) = e$. The associated **masked spectral bound** is

$$\xi_{C,s}(X) := \sum_{i=1}^{s} \ln (\lambda_i (C \circ X))$$

**Special combinatorial cases:**
- Spectral bound $X := E$
- Diagonal bound $X := I$
- Spectral partition bound $X := \text{Diag}_i(E_i)$
Validity

Based on

- \( \det A = \prod_l \lambda_l(A) \)
- “Oppenheim’s Inequality”

\[
\det A \leq \det A \circ B / \prod_{j=1}^n B_{jj} ,
\]

where \( A \succeq 0 \) and \( B \succeq 0 \)

- the eigenvalue inequalities \( \lambda_l(A) \geq \lambda_l(A') \), where \( A \succeq 0 \), and \( A' \) is a principal submatrix of \( A \)
Some References