Active Learning, Experimental Design

CS294 Practical Machine Learning

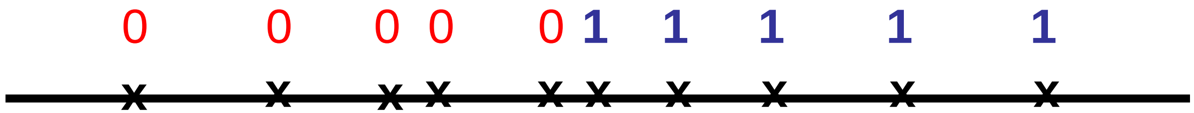
Daniel Ting

Original Slides by Barbara Engelhardt and Alex Shyr

Motivation

- Better data is often more useful than simply more data (quality over quantity)
- Data collection may be expensive
  - Cost of time and materials for an experiment
  - Cheap vs. expensive data
    - Raw images vs. annotated images
- Want to collect best data at minimal cost

Toy Example: 1D classifier



Unlabeled data: labels are all 0 then all 1 (left to right)

Classifier (threshold function): $h_w(x) = 1$ if $x > w$ (0 otherwise)

Goal: find transition between 0 and 1 labels in minimum steps

Naïve method: choose points to label at random on line

- Requires O(n) training data to find underlying classifier

Better method: binary search for transition between *0* and *1*

- Requires O(log n) training data to find underlying classifier
- Exponential reduction in training data size!

Example: collaborative filtering

- Users usually rate only a few movies; ratings “expensive”
- Which movies do you show users to best extrapolate movie preferences?
    - Also known as *questionnaire design*
- Baseline questionnaires:
    – Random: *m* movies randomly
    – Most Popular Movies: *m* most frequently rated movies
- Most popular movies is **not** better than random design!
- Popular movies rated highly by all users; do not discriminate tastes

[Yu et al. 2006]

Example: Sequencing genomes

- What genome should be sequenced next?
- Criteria for selection?
- Optimal species to detect phenomena of interest

[McAuliffe et al., 2004]

Example: Improving cell culture conditions

- Grow cell culture in bioreactor
  - Concentrations of various things
    - Glucose, Lactate, Ammonia, Asparagine, etc.
  - Temperature, etc.

- Task: Find optimal growing conditions for a cell culture
- Optimal: Perform as few time consuming experiments as possible to find the optimal conditions.

Topics for today

- Introduction: Information theory
- Active learning
  - Query by committee
  - Uncertainty sampling
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
- Non-linear optimal experimental design
  - Sequential experimental design
  - Bayesian experimental design
  - Maximin experimental design
- Summary

Topics for today

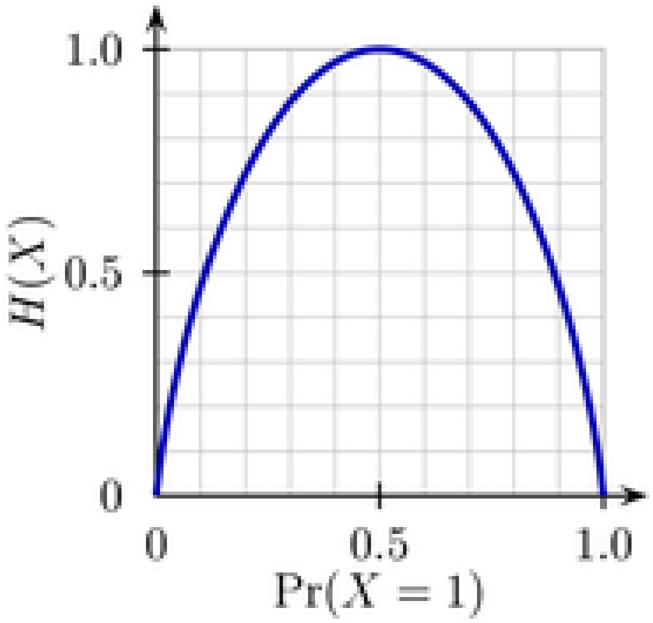
- Introduction: Information theory
- Active learning
  - Query by committee
  - Uncertainty sampling
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
- Non-linear optimal experimental design
  - Sequential experimental design
  - Bayesian experimental design
  - Maximin experimental design
- Summary

Entropy Function

- A measure of information in random event *X* with possible outcomes $\{x_1, \ldots, x_n\}$

  $$H(x) = - \Sigma_i \, p(x_i) \log_2 p(x_i)$$

- Comments on entropy function:
  - Entropy of an event is zero when the outcome is known
  - Entropy is maximal when all outcomes are equally likely
- The average minimum number of yes/no questions to answer some question
  - Related to binary search



[Shannon, 1948]

Kullback Leibler divergence

- P = true distribution;
- Q = alternative distribution that is used to encode data
- KL divergence is the expected extra message length per datum that must be transmitted using Q

$$
\begin{aligned}
D_{KL}(P \,||\, Q) &= \Sigma_i\, P(x_i) \log (P(x_i)/Q(x_i)) \\
&= \Sigma_i\, P(x_i) \log P(x_i) - \Sigma_i\, P(x_i) \log Q(x_i) \\
&= H(P,Q) - H(P) \\
&= \text{Cross-entropy - entropy}
\end{aligned}
$$

- Measures how different the two distributions are

KL divergence properties

- Non-negative: $D(P||Q) \geq 0$

- Divergence 0 if and only if P and Q are equal:
  - $D(P||Q) = 0$ iff $P = Q$

- Non-symmetric: $D(P||Q) \neq D(Q||P)$

- Does not satisfy triangle inequality
  - $D(P||Q) \not\leq D(P||R) + D(R||Q)$

KL divergence properties

- Non-negative: $D(P||Q) \geq 0$

- Divergence 0 if and only if P and Q are equal:
  - $D(P||Q) = 0 \text{ iff } P = Q$

- Non-symmetric: $D(P||Q) \neq D(Q||P)$

- Does not satisfy triangle inequality
  - $\mathrm{D(P||Q)} \not\leq \mathrm{D(P||R) + D(R||Q)}$

Not a distance metric

KL divergence as gain

- Modeling the KL divergence of the posteriors measures the amount of information gain expected from query (where x' is the queried data)*:*

$$D(\, p(\theta \mid x, x') \,||\, p(\theta \mid x))$$

- Goal: choose a query that *maximizes* the KL divergence between posterior and prior
- Basic idea: largest KL divergence between updated posterior probability and the current posterior probability represents largest gain

Topics for today

- Introduction: information theory
- Active learning
  - Query by committee
  - Uncertainty sampling
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
- Non-linear optimal experimental design
  - Sequential experimental design
  - Bayesian experimental design
  - Maximin experimental design
- Summary

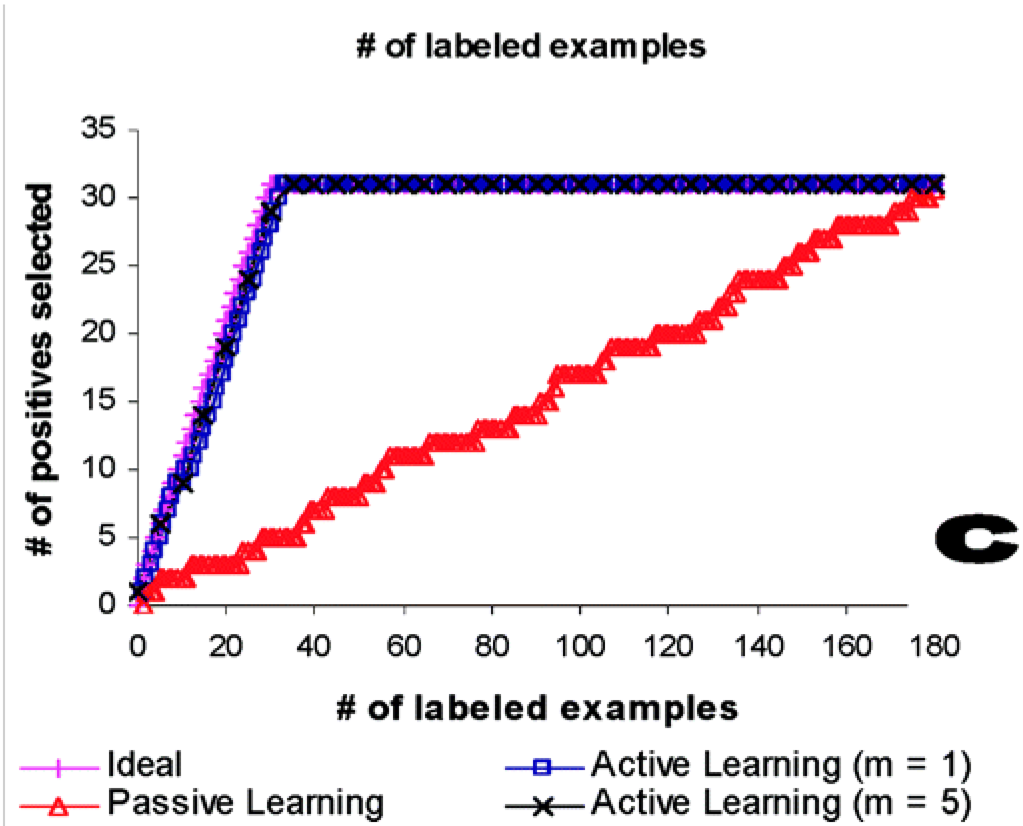
Active learning

- Setup: Given existing knowledge, want to choose where to collect more data
  - Access to cheap unlabelled points
  - Make a query to obtain expensive label
  - Want to find labels that are "informative"
- Output: Classifier / predictor trained on less labeled data

- Similar to "active learning" in classrooms
  - Students ask questions, receive a response, and ask further questions
  - vs. passive learning: student just listens to lecturer
- This lecture covers:
  - how to measure the value of data
  - algorithms to choose the data

Example: Gene expression and Cancer classification

- Active learning takes 31 points to achieve same accuracy as passive learning with 174



Liu 2004

Reminder: Risk Function

- Given an estimation procedure / decision function d
- Frequentist risk given the true parameter θ is expected loss after seeing new data.

$$R(\theta, d) = \sum_{\theta} L(\theta, d(x_{new}))p(x_{new}|\theta)$$

- Bayesian integrated risk given a prior π is defined as posterior expected loss:

$$R(\pi, d|x) = \sum_{\theta} L(\theta, d(x))p(\theta|x, \pi)$$

- Loss includes cost of query, prediction error, etc.

Decision theoretic setup

- Active learner
  - Decision d includes which data point q to query
    - also includes prediction / estimate / etc.
  - Receives a response from an oracle
- Response updates parameters θ of the model
- Make next decision as to which point to query based on new parameters

- Query selected should minimize risk

$$\min_{query} R(\theta, query)$$

Active Learning

- Some computational considerations:
  - May be many queries to calculate risk for
    - Subsample points
    - Probability far from the true min decreases exponentially
  - May not be easy to calculate risk R

- Two heuristic methods for reducing risk:
  - Select “most uncertain” data point given model and parameters
  - Select “most informative” data point to optimize expected gain

Uncertainty Sampling

- Query the event that the current classifier is most uncertain about

- Needs measure of uncertainty, probabilistic model for prediction
- Examples:
  – Entropy
  – Least confident predicted label

$$x^* = \arg\min_x P(\hat{y}|x,\theta) = \arg\min_x \max_y P(y|x,\theta)$$

  – Euclidean distance (e.g. point closest to margin in SVM)

Example: Gene expression and Cancer classification

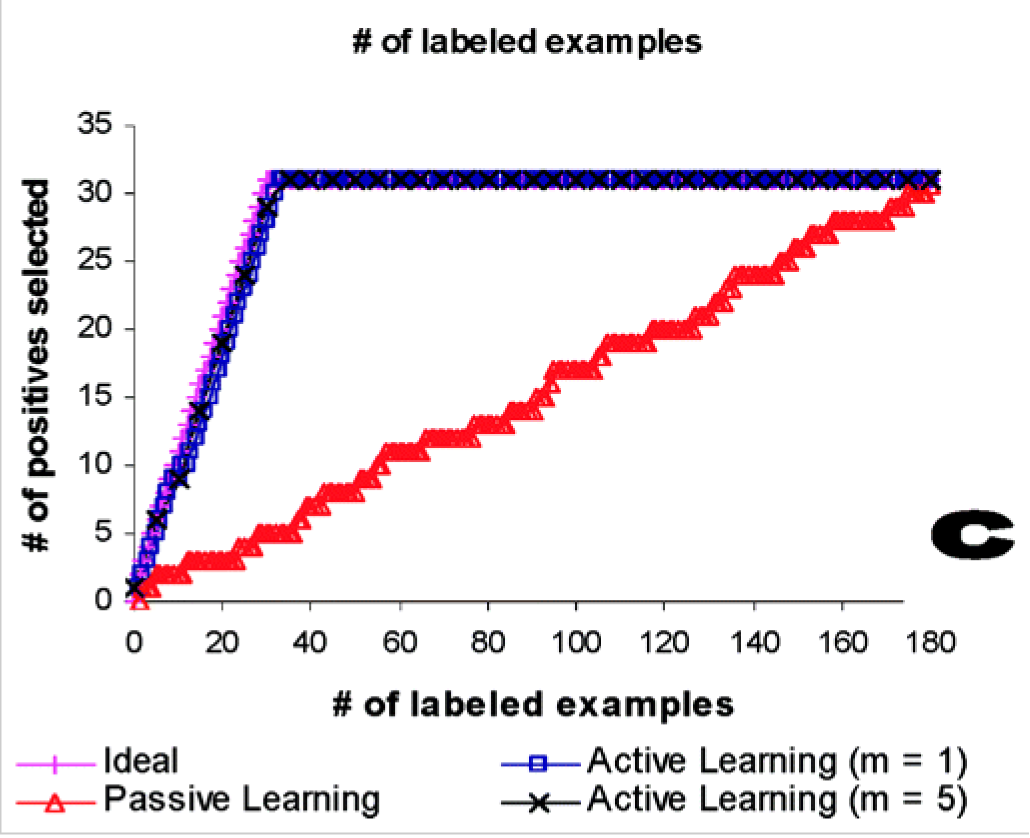
- Data: Cancerous Lung tissue samples
  - “Cheap” unlabelled data
    - gene expression profiles from Affymatrix microarray
  - Labeled data:
    - 0-1 label for adenocarcinoma or malignant pleural mesothelioma

- Method:
  - Linear SVM
  - Measure of uncertainty
    - distance to SVM hyperplane

Liu 2004

Example: Gene expression and Cancer classification

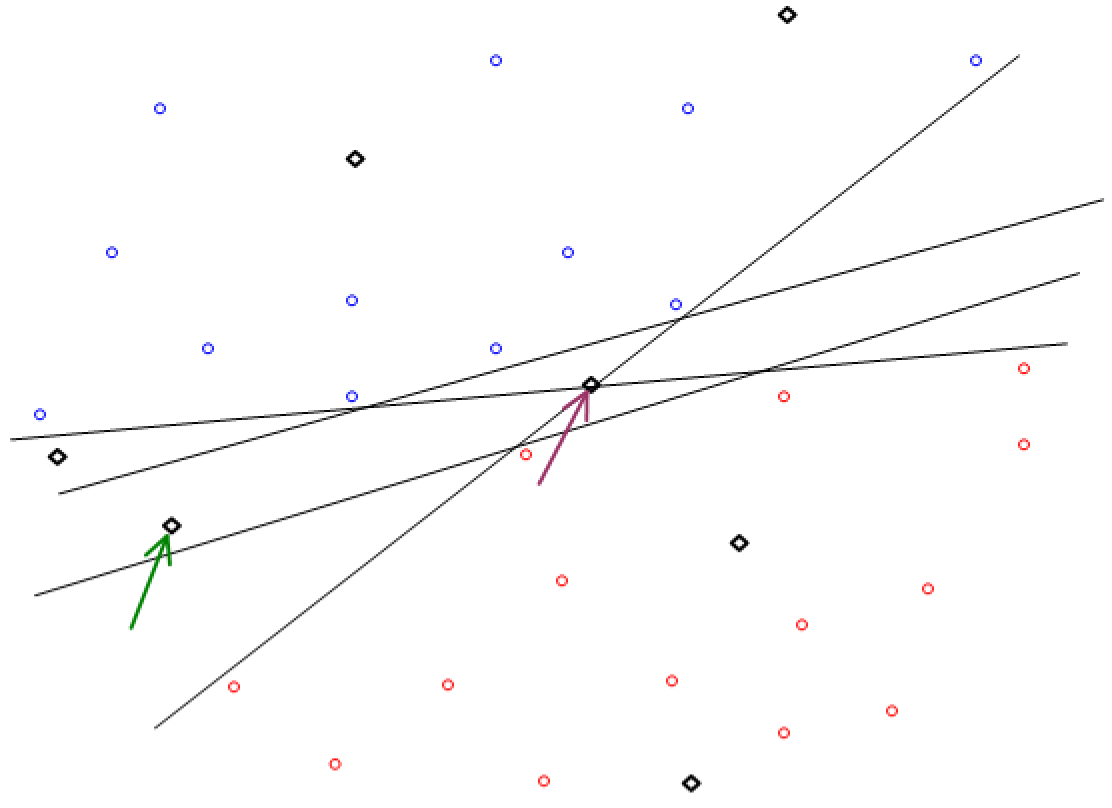
- Active learning takes 31 points to achieve same accuracy as passive learning with 174



Liu 2004

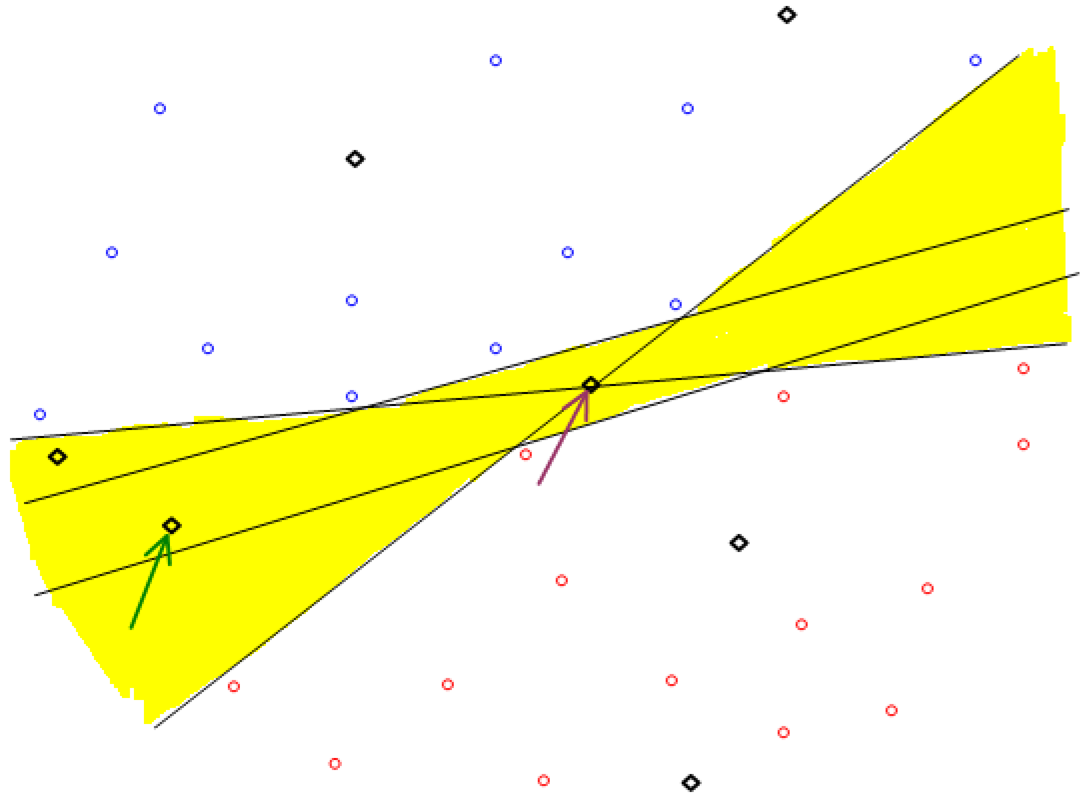
Query by Committee

- Which unlabelled point should you choose?



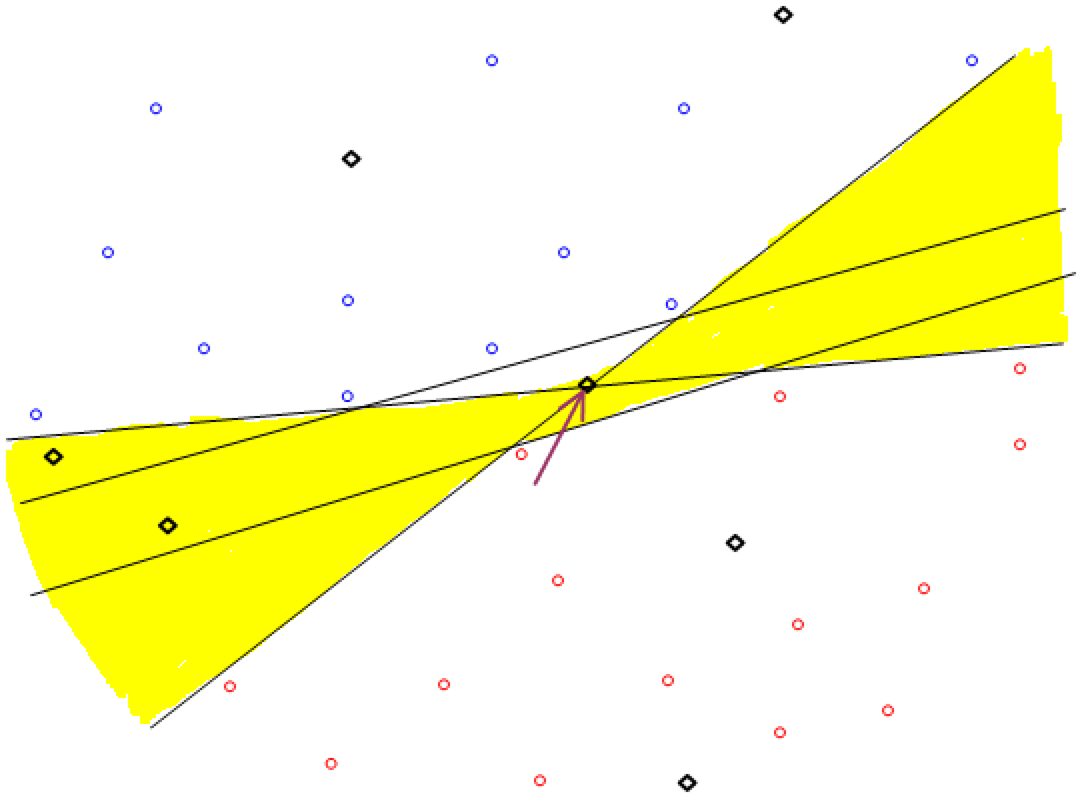
Query by Committee

- Yellow = valid hypotheses



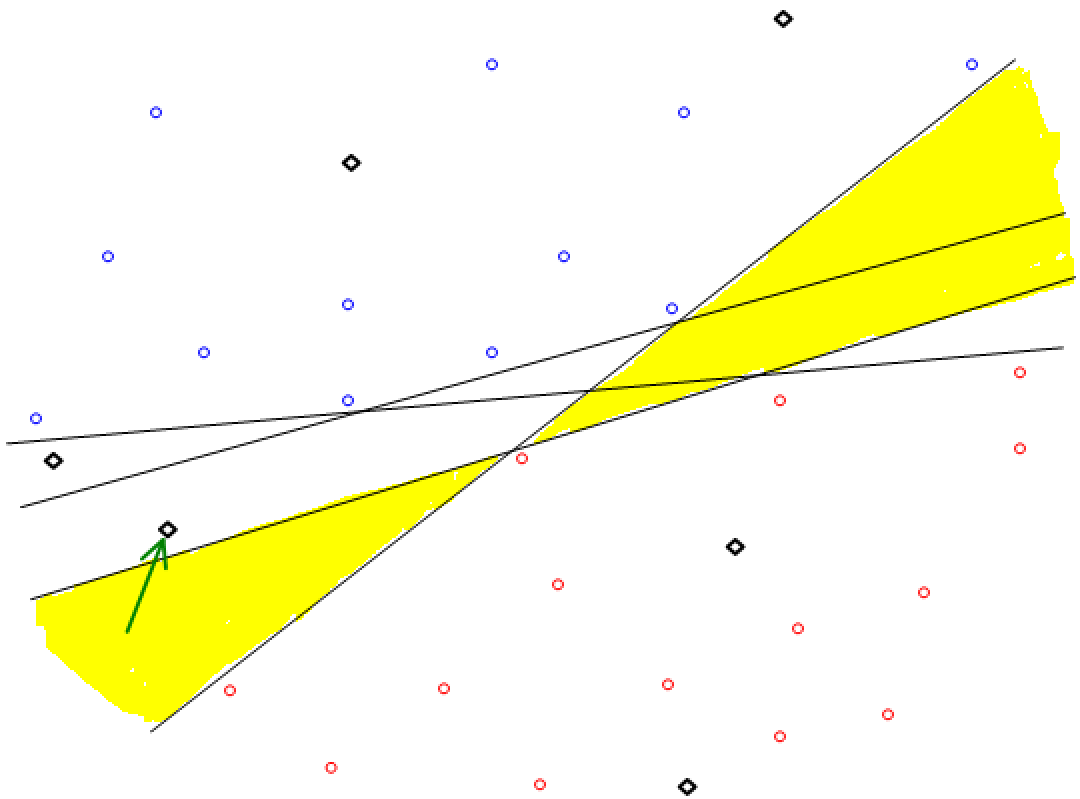
Query by Committee

- Point on max-margin hyperplane does not reduce the number of valid hypotheses by much



Query by Committee

- Queries an example based on the degree of disagreement between committee of classifiers



Query by Committee

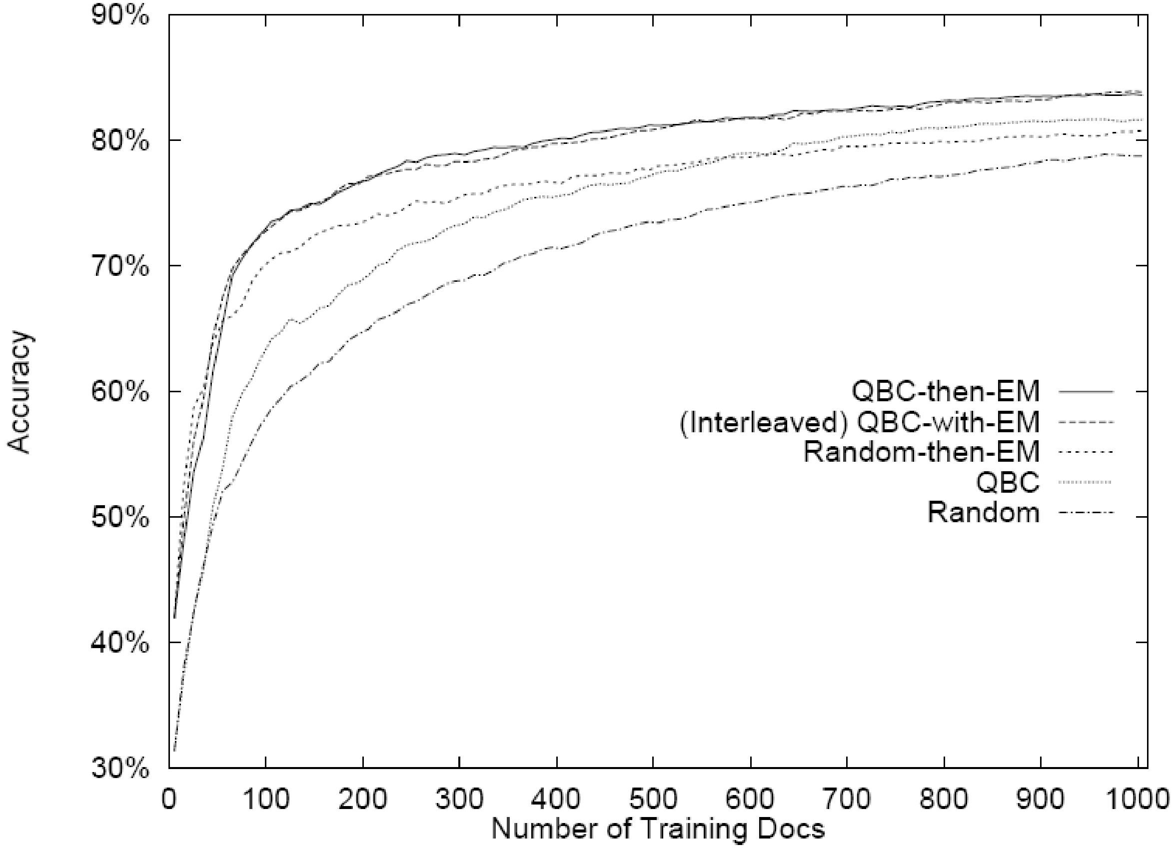
- Prior distribution over classifiers/hypotheses
- Sample a set of classifiers from distribution

- Natural for ensemble methods which are already samples
  – Random forests, Bagged classifiers, etc.

- Measures of disagreement
  – Entropy of predicted responses
  – KL-divergence of predictive distributions

Query by Committee Application

- Used naïve Bayes model for text classification in a Bayesian learning setting (20 Newsgroups dataset)



[McCallum & Nigam, 1998]

Information-based Loss Function

- Previous methods looked at uncertainty at a single point
  - Does not look at whether you can actually reduce uncertainty or if adding the point makes a difference in the model
- Want to model notions of information gained
  - Maximize **KL divergence** between posterior and prior

$$KL(P\|\pi) = \#\text{ of bits gained about model}$$

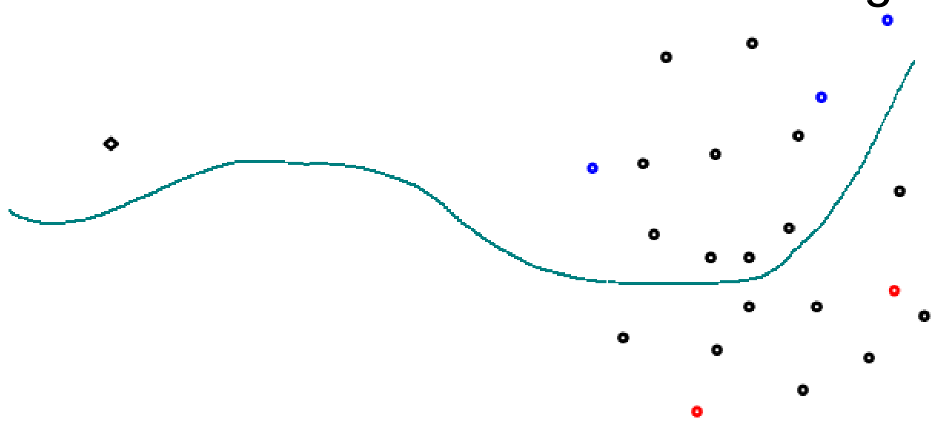
  - Maximize reduction in **model entropy** between posterior and prior (reduce number of bits required to describe distribution)

- All of these can be extended to optimal design algorithms
- Must decide how to handle uncertainty about query response, model parameters

[MacKay, 1992]

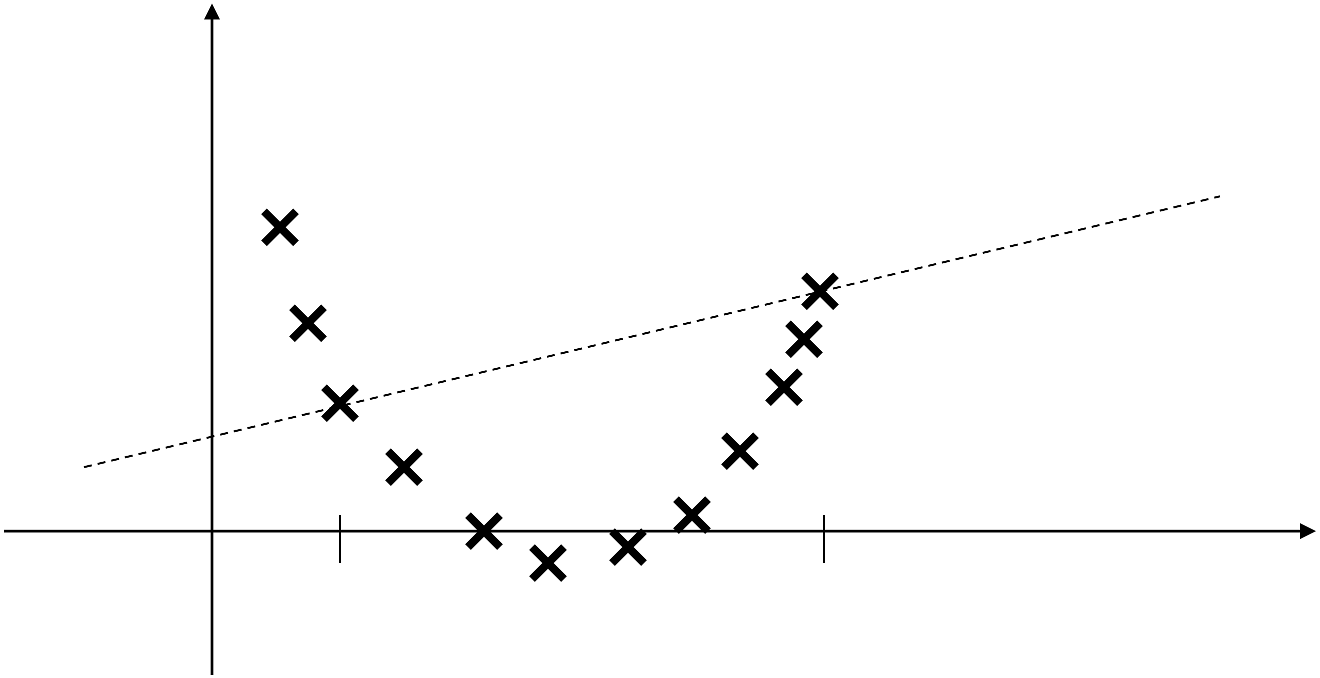
Other active learning strategies

- Expected model change
  - Choose data point that imparts greatest change to model
- Variance reduction / Fisher Information maximization
  - Choose data point that minimizes error in parameter estimation
  - Will say more in design of experiments
- Density weighted methods
  - Previous strategies use query point and distribution over models
  - Take into account data distribution in surrogate for risk.



Active learning warning

- Choice of data is only as good as the model itself
- Assume a linear model, then two data points are sufficient
- What happens when data are not linear?



Break?

Topics for today

- Introduction: information theory
- Active learning
  - Query by committee
  - Uncertainty sampling
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
- Non-linear optimal experimental design
  - Sequential experimental design
  - Bayesian experimental design
  - Maximin experimental design
- Summary

Experimental Design

- Many considerations in designing an experiment
  - Dealing with confounders
  - Feasibility
  - Choice of variables to measure
  - Size of experiment (# of data points)
  - Conduction of experiment
  - Choice of interventions/queries to make
  - Etc.

Experimental Design

- Many considerations in designing an experiment
  - Dealing with confounders
  - Feasibility
  - Choice of variables to measure
  - Size of experiment (# of data points)
  - Conduction of experiment
  - Choice of interventions/queries to make
  - Etc.
- We will only look at one of them

What is optimal experimental design?

- Previous slides give
  - General formal definition of the problem to be solved (which may be not tractable or not worth the effort)
  - heuristics to choose data
- Empirically good performance but
  - Not that much theory on how good the heuristics are

- Optimal experimental design gives
  - theoretical credence to choosing a set of points
  - for a specific set of assumptions and objectives
- Theory is good when you only get to run (a series of) experiments once

Optimal Experimental Design

- Given a model *M* with parameters β,
  - What queries are maximally informative
    i.e. will yield the best estimate of β

- "Best" minimizes variance of estimate $\hat{\beta}$
  - Equivalently, maximizes the Fisher Information

$$I(\beta) \approx var(\hat{\beta})^{-1} \quad \text{if } \hat{\beta} \text{ is the mle}$$

- Linear models
  - Optimal design does not depend on β !
- Non-linear models
  - Depends on β, but can Taylor expand to linear model

Optimal Experimental Design

- Assumptions
  - Linear model: $Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i$
  - Finite set of queries $\{F_1, \ldots, F_s\}$ that $x_{.j}$ can take.
    - Each F_i is set of interventions/measurements (e.g. F_1 =10ml of dopamine on mouse with mutant gene G)
    - m_i = # responses for query F_i
  - Usual assumptions for linear least squares regression

$$E\epsilon_i = 0 \qquad \text{(Unbiased)}$$

$$Var(\epsilon_i) = \sigma^2 \qquad \text{(Constant variance/Homoskedastic)}$$

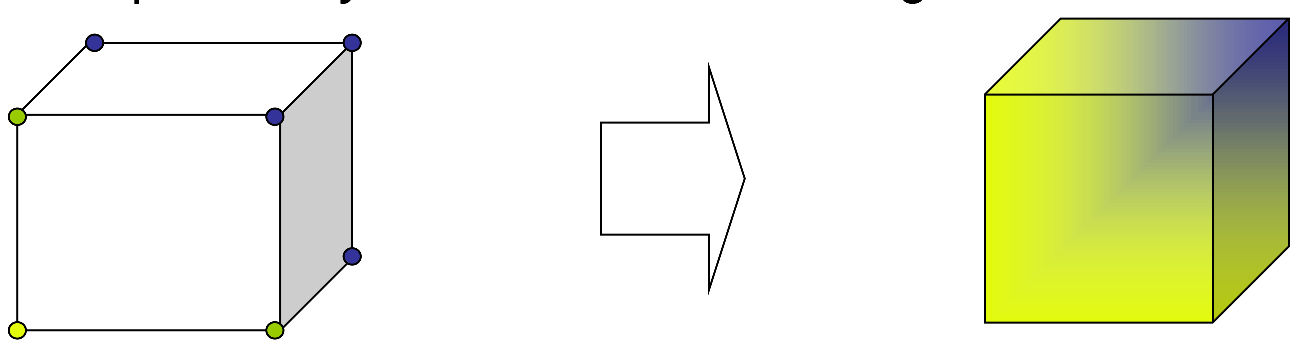
$$E\epsilon_i\epsilon_j = 0 \qquad \text{(Uncorrelated)}$$

- Covariance of mle: $Var(\hat{\beta}) = (F^T M F)^{-1}$

Relaxed Experimental Design

- Hard combinatorial problem $(F^T M F)^{-1}$
- The *relaxed* problem allows $w_i \geq 0,\ \Sigma_i w_i = 1$
- Error covariance matrix becomes $(F^T W F)^{-1}$

- $(F^T W F)^{-1}$ = inverted Hessian of the squared error
  - or inverted Fisher information matrix
- minimizing $(F^T W F)^{-1}$ reduces model error,
  - or equivalently maximize information gain



Boolean problem

N = 3

Relaxed problem

Experimental Design: Types

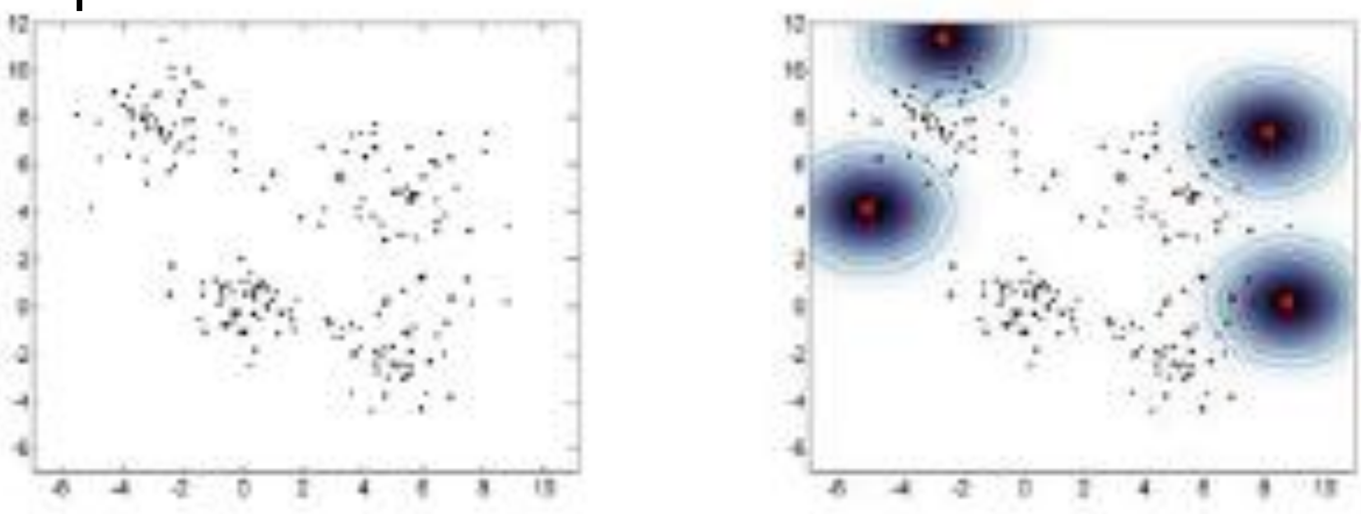
- Want to minimize $(F^T WF)^{-1}$; need a scalar objective
  - *A-optimal* (average) design minimizes trace$(F^T WF)^{-1}$
  - *D-optimal* (determinant) design minimizes log det$(F^T WF)^{-1}$
  - *E-optimal* (extreme) design minimizes max eigenvalue of $(F^T WF)^{-1}$
  - Alphabet soup of other criteria (C-, G-, L-, V-,etc)
- All of these design methods can use convex optimization techniques
- Computational complexity polynomial for semi-definite programs (*A-* and *E-optimal* designs)

[Boyd & Vandenberghe, 2004]

A-Optimal Design

- *A-optimal* design minimizes the trace of $(F^T W F)^{-1}$
  - Minimizing trace (sum of diagonal elements) essentially chooses maximally independent columns (small correlations between interventions)
- Tends to choose points on the border of the dataset

Example: mixture of four Gaussians



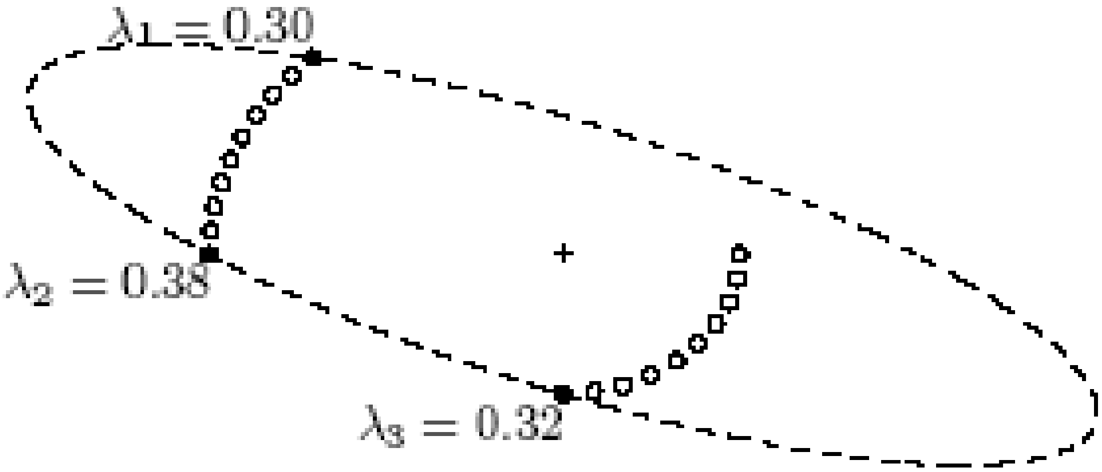
(a) Data set (b) A-optimal design

[Yu et al., 2006]

A-Optimal Design

- *A-optimal* design minimizes the trace of $(F^T W F)^{-1}$
  - Can be cast as a semi-definite program

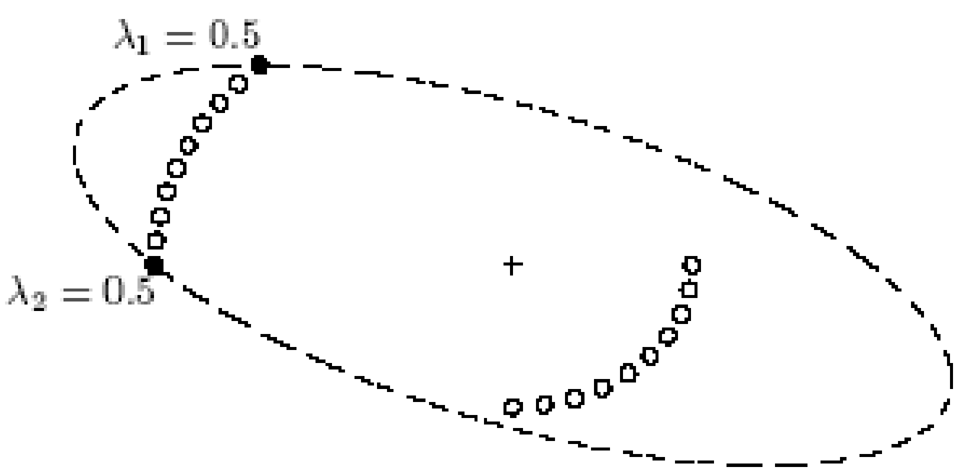
Example: 20 candidate datapoints, minimal ellipsoid that contains all points



[Boyd & Vandenberghe, 2004]

D-Optimal design

- *D-optimal* design minimizes log determinant of $(F^T W F)^{-1}$
- Equivalent to
  - choosing the confidence ellipsoid with minimum volume ("most powerful" hypothesis test in some sense)
  - Minimizing entropy of the estimated parameters $\hat{\beta}$
- Most commonly used optimal design



[Boyd & Vandenberghe, 2004]

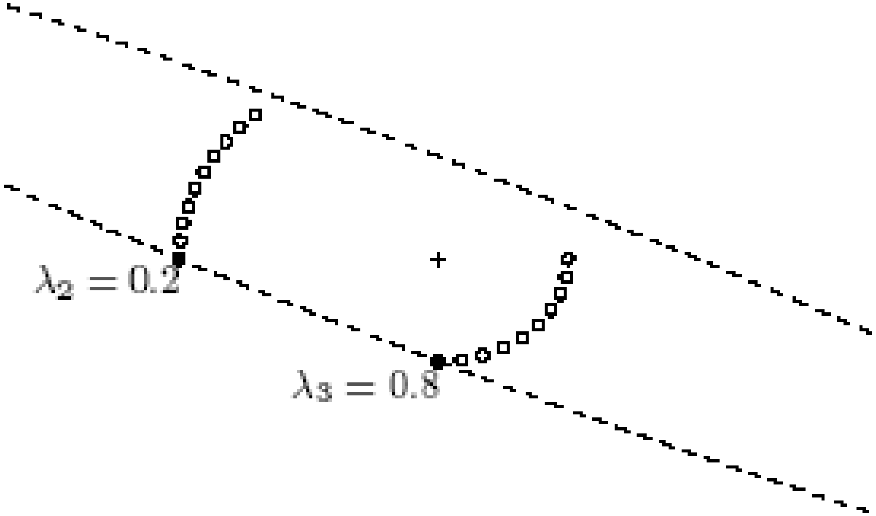
E-Optimal design

- *E-optimal* design minimizes largest eigenvalue of $(F^T W F)^{-1}$
- Minimax procedure

$$\min_W \max \; eigenvalues(F^T W F)^{-1}$$

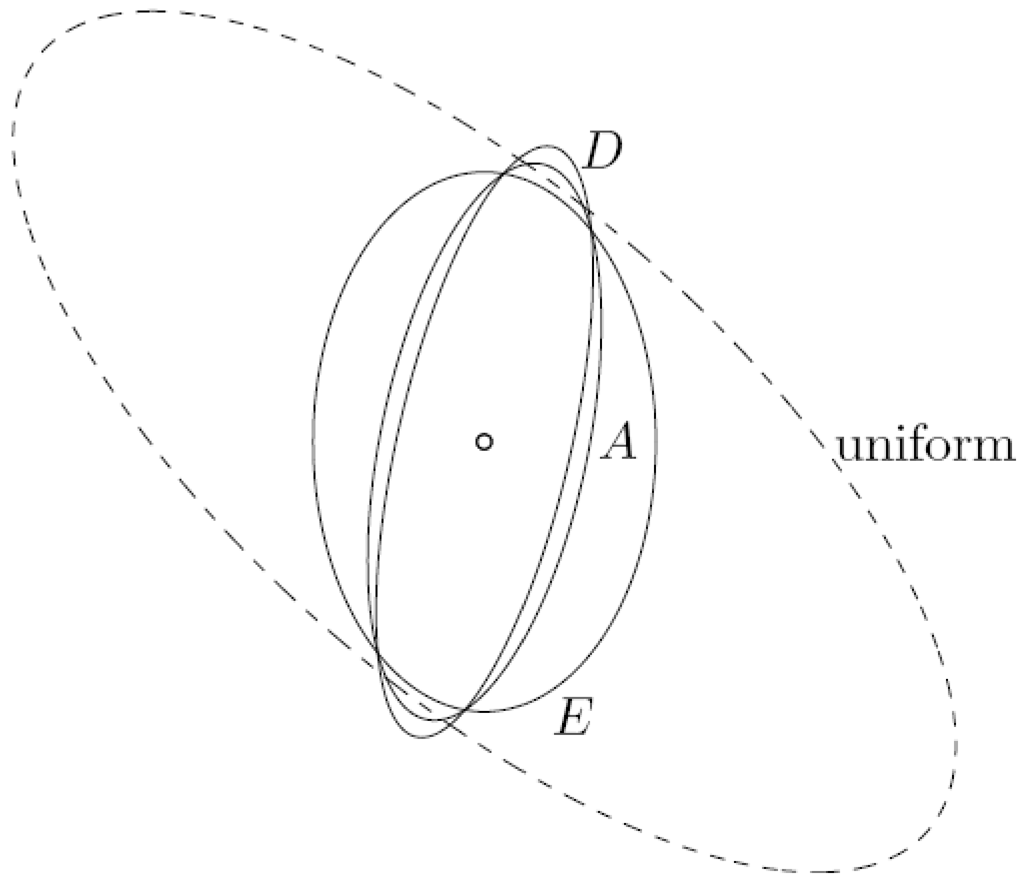
- Can be cast as a semi-definite program

- Minimizes the diameter of the confidence ellipsoid



[Boyd & Vandenberghe, 2004]

Summary of Optimal Design



[Boyd & Vandenberghe, 2004]

Optimal Design

- Extract the integral solution from the relaxed problem

- Can simply round the weights to closest multiple of 1/m

  – $m_j = \mathrm{round}(m * w_i)$, $i = 1, \ldots, p$

[Boyd & Vandenberghe, 2004]

Extensions to optimal design

- Cost associated with each experiment
  – Add a cost vector, constrain total cost by a budget B (one additional constraint)

- Multiple samples from single experiment
  – Each x_i is now a matrix instead of a vector
  – Optimization (covariance matrix) is identical to before

- Time profile of process
  – Add time dimension to each experiment vector x_i

[Atkinson, 1996]
[Boyd & Vandenberghe, 2004]

Topics for today

- Introduction: information theory
- Active learning
  - Query by committee
  - Uncertainty sampling
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
- Non-linear optimal experimental design
  - Sequential experimental design
  - Bayesian experimental design
  - Maximin experimental design
- Summary

Optimal design in non-linear models

- Given a non-linear model $y = g(x,\theta)$
- Model is described by a Taylor expansion around a $\hat{\theta}$
  - $a_j(x, \hat{\theta}) = \partial g(x,\theta) / \partial \theta_j$, evaluated at $\hat{\theta}$

$$Y_i = g(x,\hat{\theta}) + (\theta_1 - \hat{\theta}_1)a_1(x,\hat{\theta}) + ... + (\theta_k - \hat{\theta}_k)a_k(x,\hat{\theta})$$

- Maximization of Fisher information matrix is now the same as the linear model

- Yields a locally optimal design, optimal for the particular value of θ
- Yields no information on the (lack of) fit of the model

[Atkinson, 1996]

Optimal design in non-linear models

- *Problem*: parameter value θ, used to choose experiments F, is unknown
- Three general techniques to address this problem, useful for many possible notions of "gain"

- **Sequential experimental design**: iterate between choosing experiment x and updating parameter estimates θ
- **Bayesian experimental design**: put a prior distribution on parameter θ, choose a best data x
- **Maximin experimental design**: assume worst case scenario for parameter θ, choose a best data x

Sequential Experimental Design

- Model parameter values are not known exactly
- Multiple experiments are possible
- Learner assumes that only one experiment is possible; makes best guess as to optimal data point for given θ
- Each iteration:
  - Select data point to collect via experimental design using θ
  - Single experiment performed
  - Model parameters θ' are updated based on all data x'

- Similar idea to Expectation Maximization

[Pronzato & Thierry, 2000]

Bayesian Experimental Design

- Effective when knowledge of distribution for θ is available
- Example: KL divergence between posterior and prior
  - $\int_x \text{argmax}_w \int_{\theta\in\Theta} D(\, p(\theta\,|w,x) \,||\, p(\theta)) \, p(x\,|w) \, d\theta \, dx$
- Example: A-optimal design:
  - $\int_x \text{argmin}_w \int_{\theta\in\Theta} \text{tr}(F^T W F)^{-1} p(\theta\,|\, w,x) p(x\,|w) \, d\theta \, dx$
- Often sensitive to distributions

[Chaloner & Verdinelli, 1995]

Maximin Experimental Design

- Maximize the minimum gain
- Example: D-optimal design:
  - argmax $\min_{\theta\in\Theta} I(\hat{\theta})$ = $\text{argmin}_w \max_{\theta\in\Theta}$ log det $(F^T W F)^{-1}$
- Example: KL divergence:
  - $\text{argmax}_w \min_{\theta\in\Theta} D(p(\theta|w,x) \,||\, p(\theta))$

- Does not require prior/empirical knowledge
- Good when very little is known about distribution of parameter θ

[Pronzato & Walter, 1988]

Topics for today

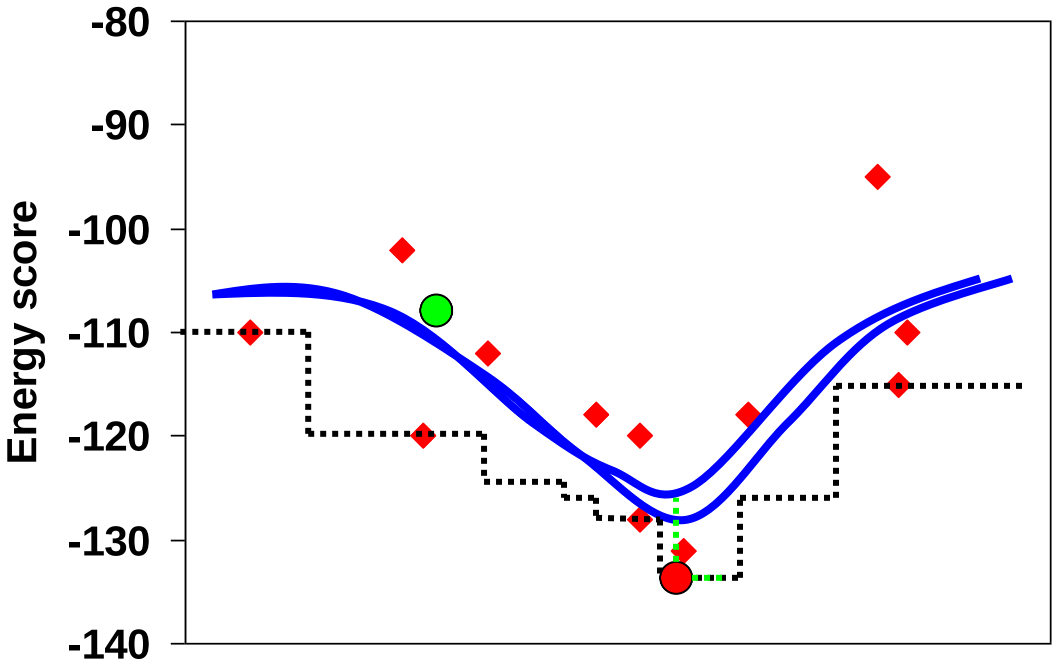
- Introduction: information theory
- Active learning
  - Query by committee
  - Uncertainty sampling
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
- Non-linear optimal experimental design
  - Sequential experimental design
  - Bayesian experimental design
  - Maximin experimental design
- Response surface models
- Summary

Response Surface Methods

- Estimate effects of local changes to the interventions (queries)
  - In particular, estimate how to maximize the response
- Applications:
  - Find optimal conditions for growing cell cultures
  - Develop robust process for chemical manufacturing
- Procedure for maximizing response
  - Given a set of datapoints, interpolate a local surface (This local surface is called the "response surface")
    - Typically use a quadratic polynomial to obtain a Hessian
  - Hill-climb or take Newton step on the response surface to find next x
  - Use next x to interpolate subsequent response surface

Response Surface Modeling

- Goal: Approximate the function f(c) = score(minimize(c))



- 1. Fit a smoothed response surface to the data points
- 2. Minimize response surface to find new candidate
- 3. Use method to find nearby local minimum of score function
- 4. Add candidate to data points
- 5. Re-fit surface, repeat

[Blum, unpublished]

Related ML Problems

- Reinforcement Learning
  - Interaction with the world
  - Notion of accumulating rewards
- Semi-supervised learning
  - Use the unlabelled data itself, not just as pool of queries
- Core sets, active sets
  - Select small dataset gives nearly same performance as full dataset. Fast computation for large scale problems

Summary

- Active learning
  - Query by committee

    Distribution over parameter;
    Probabilistic; sequential
  - Uncertainty sampling

    Predictive distribution on pt;
    Distance function; sequential
  - Information-based loss functions

    Maximize gain; sequential
- Optimal experimental
  - A-optimal design

    Minimize trace of information matrix
  - D-optimal design

    Minimize log det of information matrix
  - E-optimal design

    Minimize largest eigenvalue of information matrix
- Non-linear optimal experimental

  Multiple-shot experiments;
  Little known of parameter
  - Sequential experimental design
  - Bayesian experimental design

    Single-shot experiment;
    Some idea of parameter distribution
  - Maximin experimental design

    Single-shot experiment;
    Little known of parameter
    distribution (range known)
- Response surface methods

  Sequential experiments for optimization