Active Learning & Experimental Design

by Barbara Engelhardt and Alex Shyr

Modified by Daniel Ting

Heavily modified, of course, by **Lyle Ungar**

Strategies for choosing which points to label

Active learning: sequential, *ad hoc*

Experimental design: simultaneous, principled

Motivation

- **Labeling data is often expensive**
  - Unlabeled data is often cheap
- **Not all labels are equally useful**
- **We want to collect the "best" data at minimal cost**

What observations should one label?

Toy examples

Assume you are learning $y = ax+b$ for x on $[-1,1]$.

You can pick two x's to get y's for.

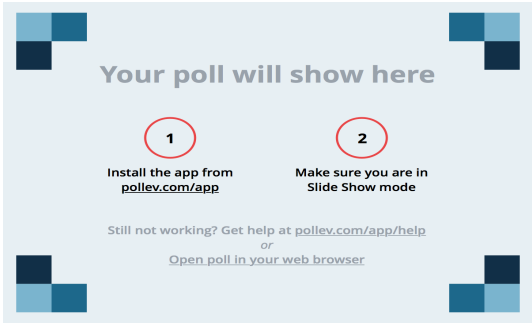
What two values would you pick?

A) -1/3, 1/3

B) -1, 1

C) 0, 1

D) Something else



Toy examples

Assume you are learning $y = f(x)$ for x a scalar

You are learning an SVM classifier on [-1.1].

You can pick 4 x's to get y's for.

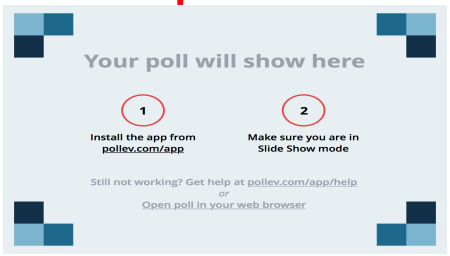
What strategy would you use to pick x's?

A) Pick -1, -1/3, 1/3, 1

B) Pick -1, 1, see what the answer is, then pick next x

C) Pick -1/3, 1/3, see what the answer is, then pick next x

D) Something else



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 are“expensive”
- **Which movies do you show users to best extrapolate movie preferences?**

[Yu et al. 2006]

Example: collaborative filtering

- **Baseline algorithms:**
  - 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]

Active Learning

- **Active learning**
  - Uncertainty sampling
  - Query by committee
  - Information-based loss functions
- **Optimal experimental design**
  - A-optimal design
  - D-optimal design
  - E-optimal design
  - Non-linear optimal experimental design

Active Learning

- **Given existing knowledge (*X*,*y*), choose where to collect more labels**
  - Assume access to cheap unlabeled points
  - Make a query to obtain expensive label
  - Want to find labels that are "informative"
  - **Output:** Classifier / predictor
- **Similar to "active learning" in classrooms**
  - Students ask questions, receive a response, and ask more questions
  - Contrast: passive learning: student just listens to lecturer

Active Learning Setup

- Active learner picks which data point *x* to query
- Receive label ("response") *y* from an oracle
- Update parameters *w* of the model
- Repeat

- Query selected to minimize some loss function ("risk")

Active Learning

- **Heuristic methods for reducing risk:**
  - Select "*most uncertain*" data point
  - Select "*most informative*" data point

Uncertainty Sampling

- **Query the item (x) that the current classifier is *most uncertain* about**
- **Needs measure of uncertainty**
- **Examples:**
  - Entropy
  - Least confident predicted label
  - Euclidean distance (e.g. point closest to margin in SVM)

When might this fail?

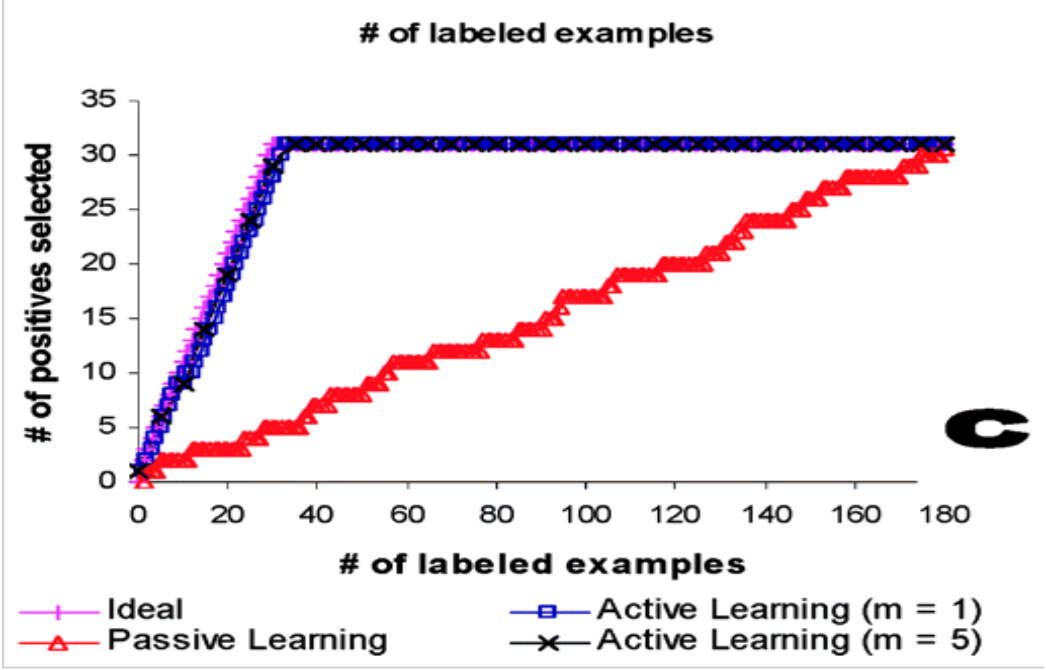
Example: Gene expression and Cancer classification

- **Data: Cancerous lung tissue samples**
  - "Cheap" unlabeled data
    - gene expression profiles from Affymetrix 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

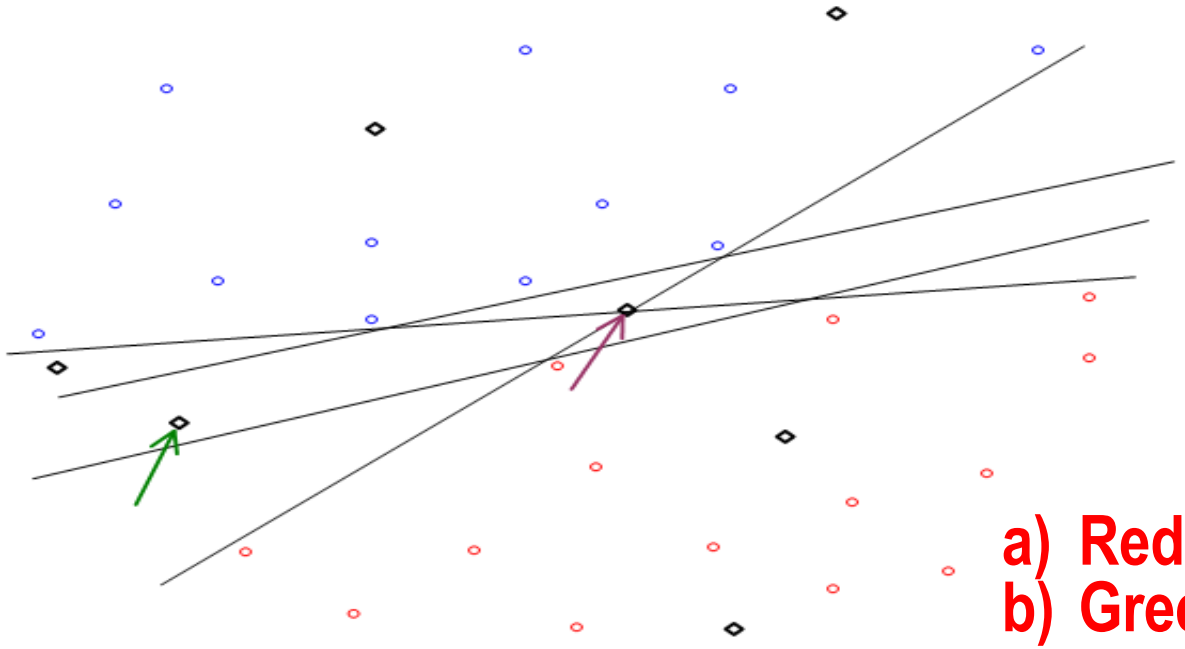
- Active learning takes 31 points to achieve same accuracy as passive learning with 174

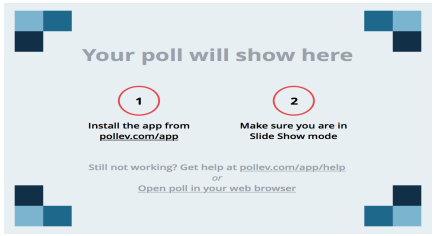


Liu 2004

Query by Committee

- Which unlabeled point should you choose?

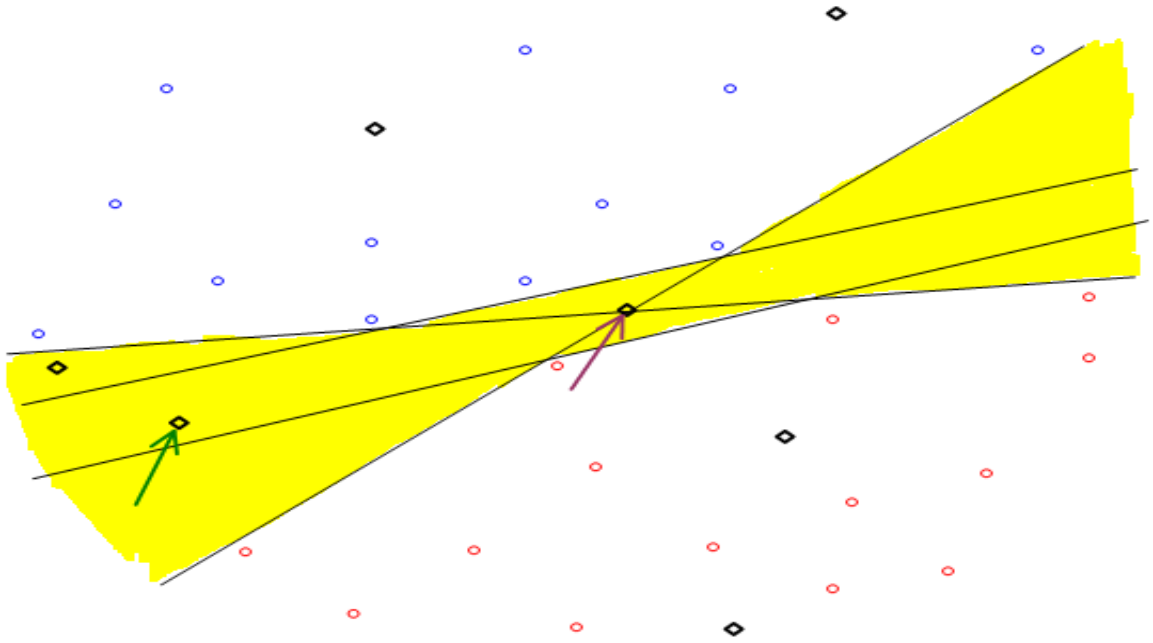




a) Red arrow point
b) Green arrow point

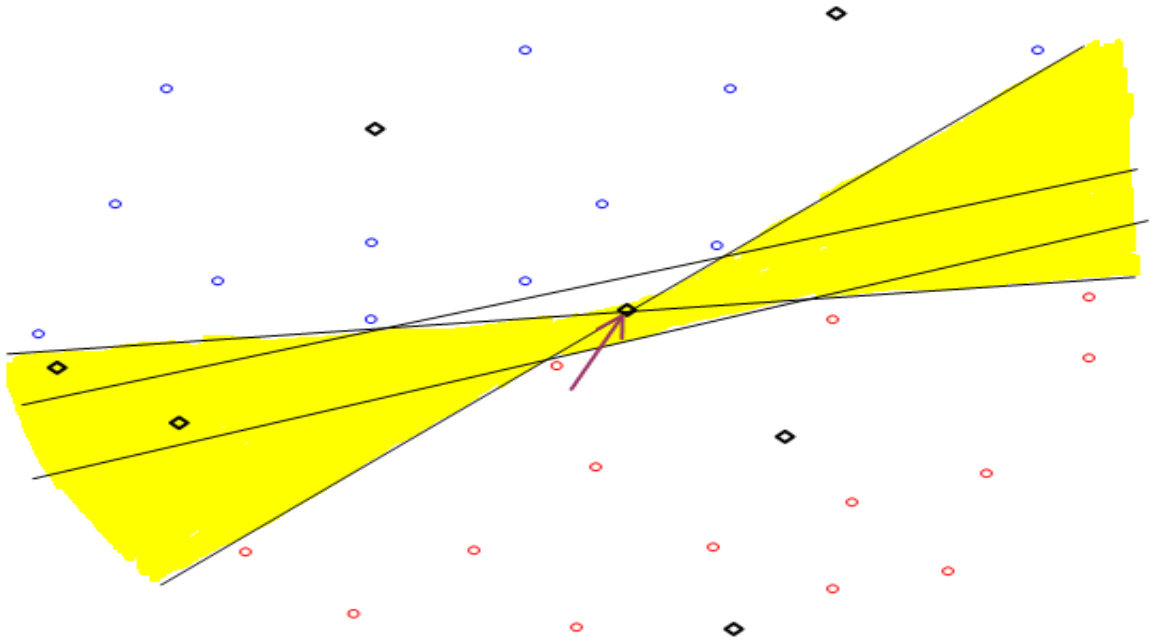
Query by Committee

- Yellow = valid hypotheses



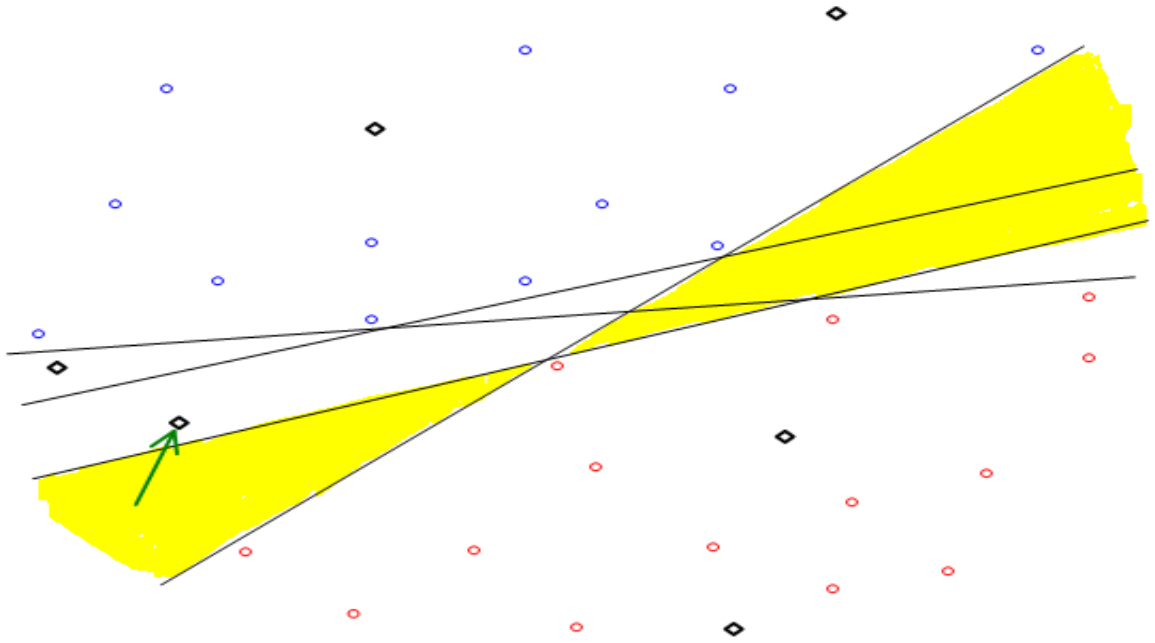
Query by Committee

- 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

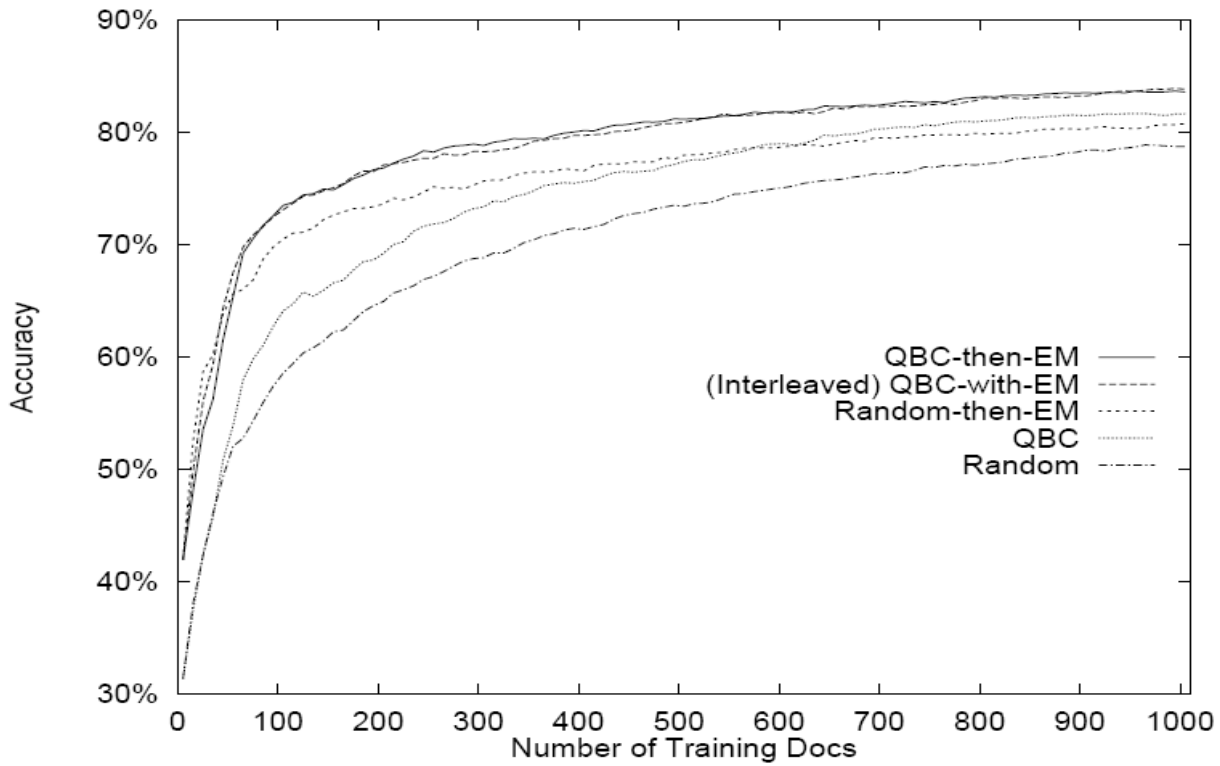


Query by Committee

- **Start with prior distribution over classifiers/hypotheses**
- **Sample a set of classifiers from distribution**
- **Natural for ensemble methods**
  - Random forests, Bagged classifiers, etc.
- **Measures of disagreement**
  - Entropy of predicted responses
  - KL-divergence of predictive distributions

Query by Committee Application

- Use Naïve Bayes model for text classification (20 Newsgroups dataset)



[McCallum & Nigam, 1998]

Information-based Loss Function

- **Above methods looked at uncertainty at a single point**
  - Does not look at expected effect of adding the point on the model
- **Want to quantify 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]

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***

$$KL(P \,||\, Q) = \sum_i P(x_i) \log (P(x_i)/Q(x_i))$$

$$= - \sum_i P(x_i) \log Q(x_i) + \sum_i P(x_i) \log P(x_i)$$

$$= H(P,Q) \qquad - H(P)$$

= Cross-entropy - entropy

- **Measures how different the two distributions are**

KL divergence as info gain

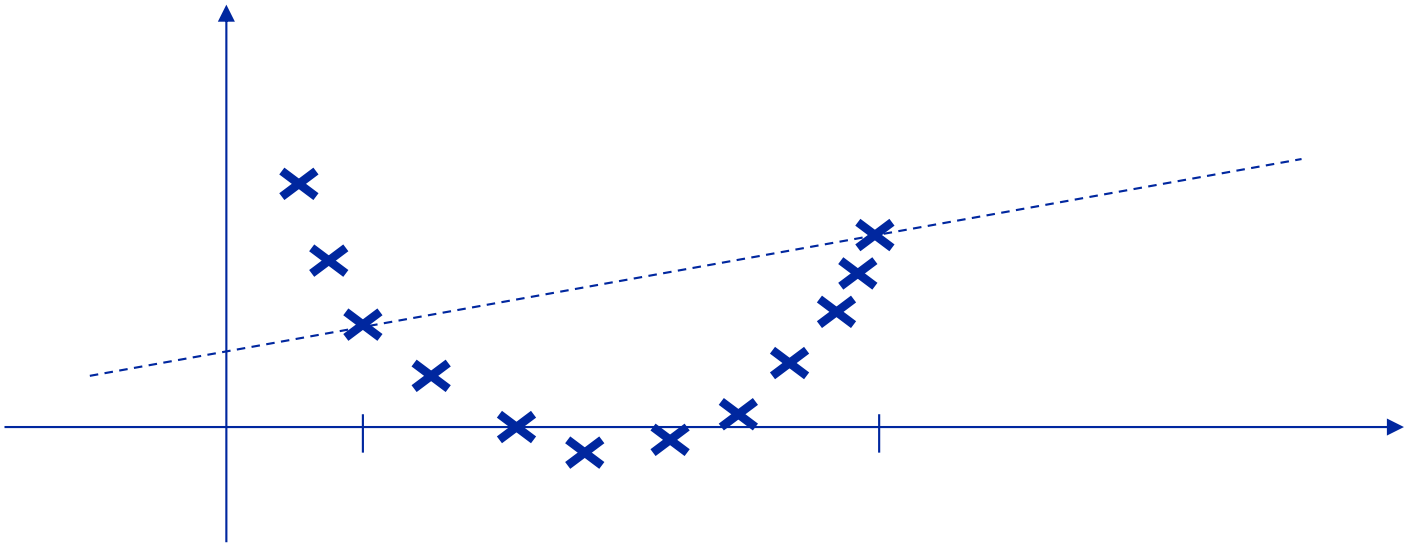
- **The KL divergence of the posteriors measures the information gain expected from query (**x'**)***:*

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

- **Goal:** choose a query that *maximizes* the KL divergence between the updated posterior probability and the current posterior probability
  - This represents the largest expected information gain

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?



Active Learning = Sequential Experimental Design

- Active learning
  - Uncertainty sampling
  - Query by committee
  - Information-based loss functions
- **Optimal experimental design**
  - A-optimal design
  - D-optimal design
  - E-optimal design
  - Non-linear optimal experimental design

Optimal Experimental Design

- **Active learning heuristics give empirically good performance but sometimes fail**

- **Optimal experimental design gives**
  - theoretical criteria for choosing a set of points to label for a specific set of assumptions and objectives

  It fails, too, if the assumptions aren't met.

Optimal Experimental Design

- **Given a model with parameters *w*,**
  - What queries are maximally informative
    i.e. will yield the best estimate of *w*
- **"Best" minimizes variance of estimate of *w***
- **Linear models**
  - Optimal design does not depend on *w* !
- **Non-linear models**
  - Depends on *w;* often use Taylor expansion to linear model

Goal: Minimize variance of *w*

If $y = x^T\beta + \varepsilon$ then $w = (X^TX)^{-1}X^Ty$

$w \sim N(\beta, \sigma^2(X^TX)^{-1})$ $\varepsilon \sim N(0, \sigma^2)$

We want to minimize the variance of our parameter estimate **w**, so pick training data **X** to minimize $(X^TX)^{-1}$

But that is a matrix, so we need to reduce it to a scalar

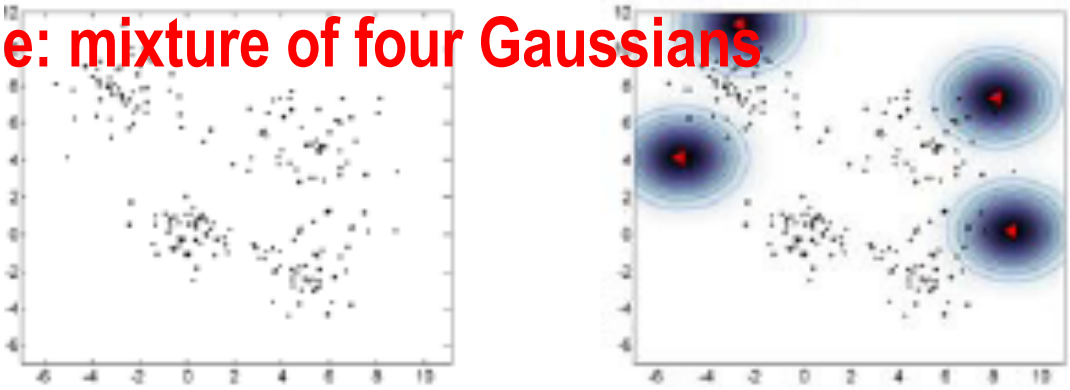
- ***A-optimal*** **(average) design minimizes** $\text{trace}(X^TX)^{-1}$
- ***D-optimal*** **(determinant) design minimizes** $\log\det(X^TX)^{-1}$
- ***E-optimal*** **(extreme) design minimizes** max eigenvalue of $(X^TX)^{-1}$

Alphabet soup of other criteria (C-, G-, L-, V-, etc.)

A-Optimal Design

- ***A-optimal* design minimizes the trace of $(X^TX)^{-1}$**
  - Minimizing trace (sum of diagonal elements) essentially chooses maximally independent columns
  - Chooses points near the border of the dataset
  - Trace of a matrix is the sum of its eigenvalues

Example: mixture of four Gaussians



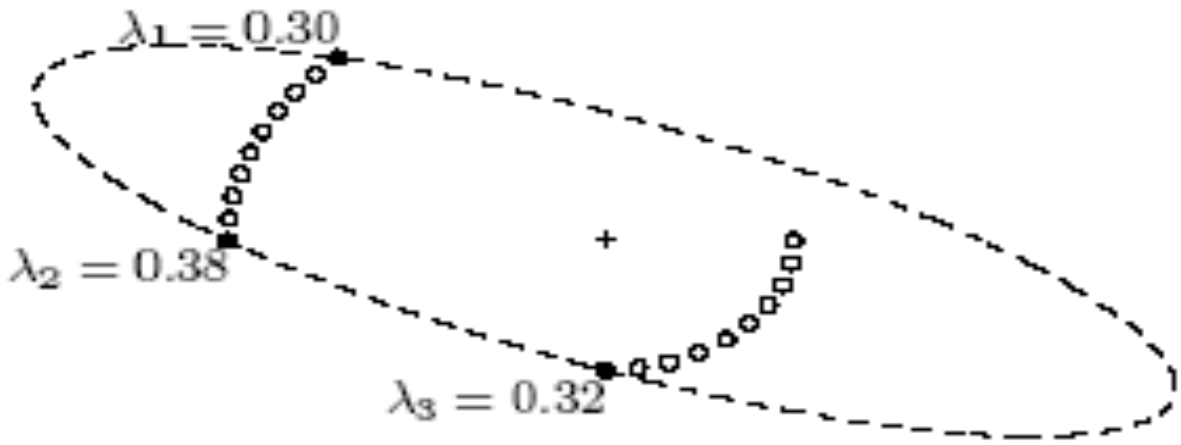
(a) Data set (b) A-optimal design

[Yu et al., 2006]

A-Optimal Design

- ***A-optimal*** **design minimizes the trace of** $(X^TX)^{-1}$

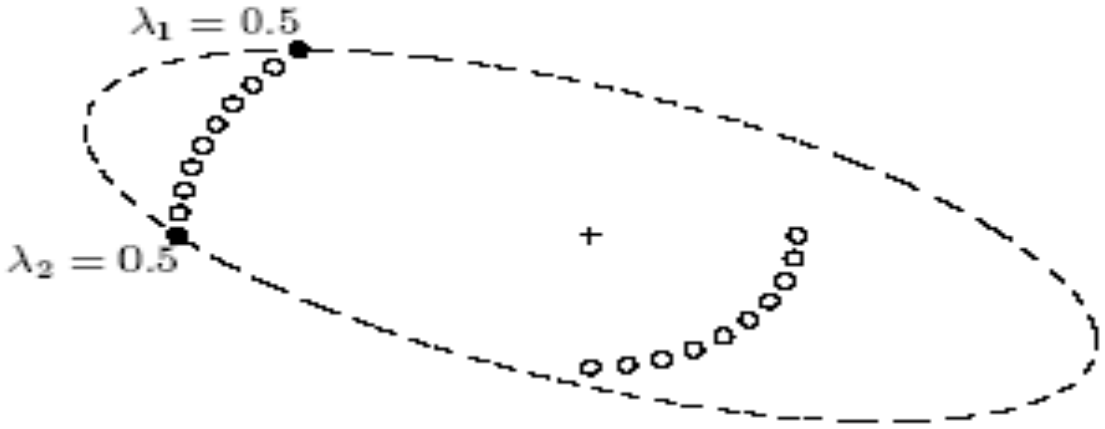
Example: 20 candidate data points, minimal ellipsoid that contains all points



[Boyd & Vandenberghe, 2004]

D-Optimal design

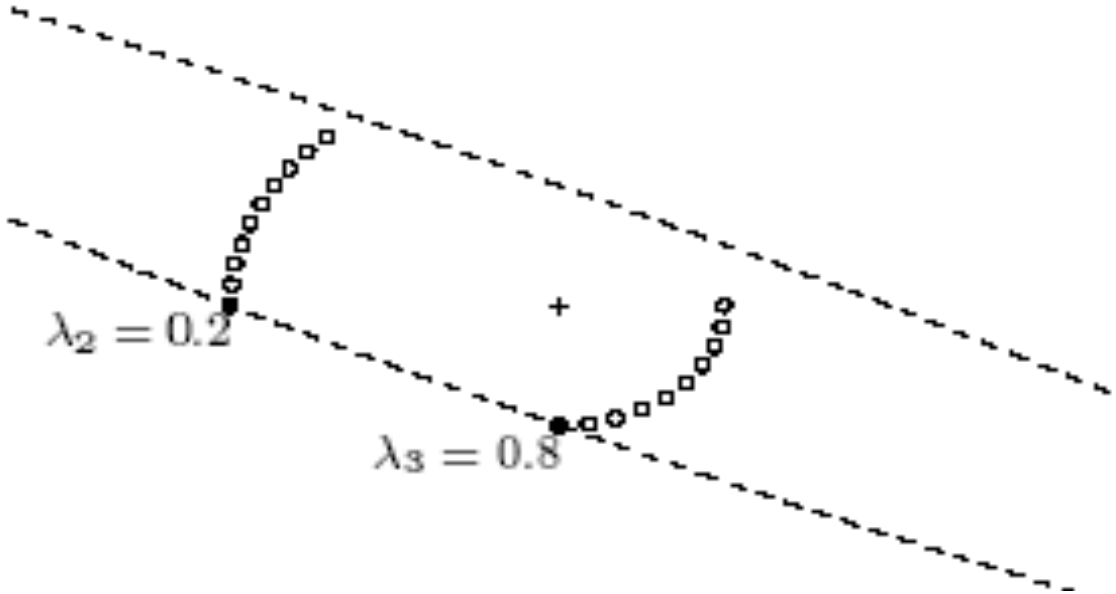
- ***D-optimal* design minimizes the determinant of $(X^TX)^{-1}$**
  - *Determinant of a matrix is the product of the eigenvalues*
  - Chooses the confidence ellipsoid with minimum volume ("most powerful" hypothesis test in some sense)
  - Minimizes entropy of the estimated parameters
- **Most commonly used optimal design**



[Boyd & Vandenberghe, 2004]

E-Optimal design

- ***E-optimal*** **design minimizes largest eigenvalue of** $(X^TX)^{-1}$
- **Minimizes the diameter of the confidence ellipsoid**



[Boyd & Vandenberghe, 2004]

Practicalities

- **Sometimes you can generate an *x* arbitrarily**
- **More often you need to select from a set of given *x*'s**
  - This can be an expensive search!

Experimental Design

- Active learning
  - Uncertainty sampling
  - Query by committee
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
  - Non-linear optimal experimental design

Optimal design in non-linear models

- **A given non-linear model *y = g(x, θ)***
- **is described by a Taylor expansion around the current estimate**
  - $a_j(x, \hat{\theta}) = \partial g(x,\theta)/\partial \theta_j$, evaluated at $\hat{\theta}$

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

- **Now just keep only the linear term**
  - so the design is the same as before
- **Yields a locally optimal design, optimal for the particular value of θ**

[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**

1) Sequential experimental design: iterate between choosing experiment *x* and updating parameter estimates θ

2) Bayesian experimental design: put a prior distribution on parameter θ, choose a best data **x**

3) Maximin experimental design: assume worst case scenario for parameter θ, choose a best data **x**

Response Surface Methods

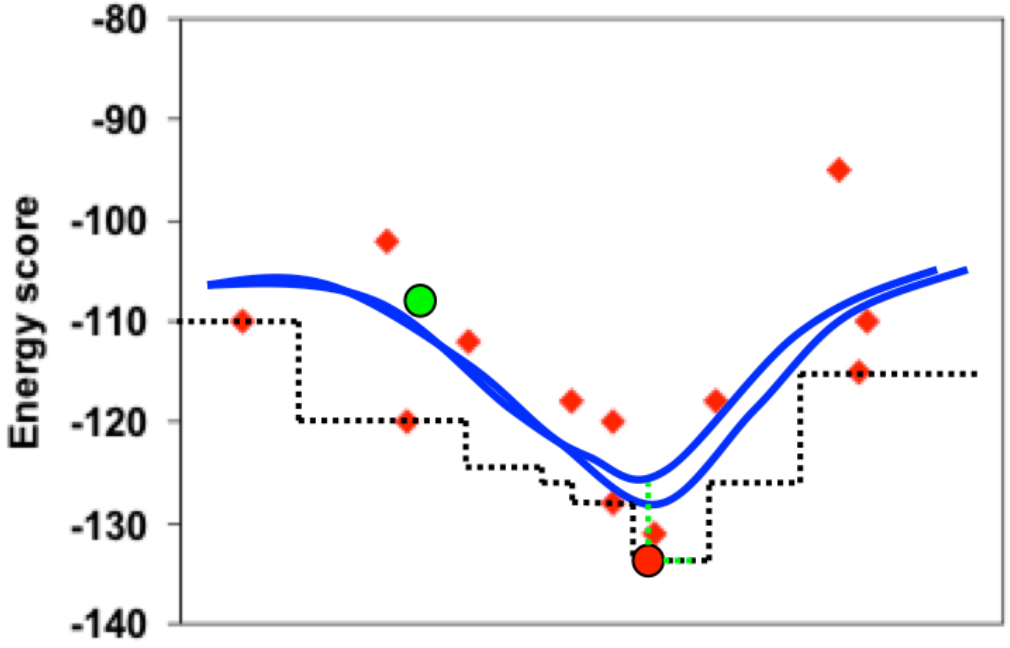
- **Estimate effects of local changes to the interventions (queries)**
  - In particular, estimate how to maximize the response
- **Applications:**
  - Find optimal conditions for growing cell cultures
  - Develop robust process for chemical manufacturing
- **Procedure for maximizing response**
  - Given 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]

Summary

- **Active learning (sequential)**
  - 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
- **Response surface methods**

Multiple models

Predictive distribution on points

Maximize info gain

Minimize trace of $(X^TX)^{-1}$

Minimize det of $(X^TX)^{-1}$

Minimize largest eigenvalue of $(X^TX)^{-1}$

Multiple-shot experiments;
Little known of parameters

Single-shot experiment;
Some idea of parameter distribution

Sequential experiments for optimization