

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?

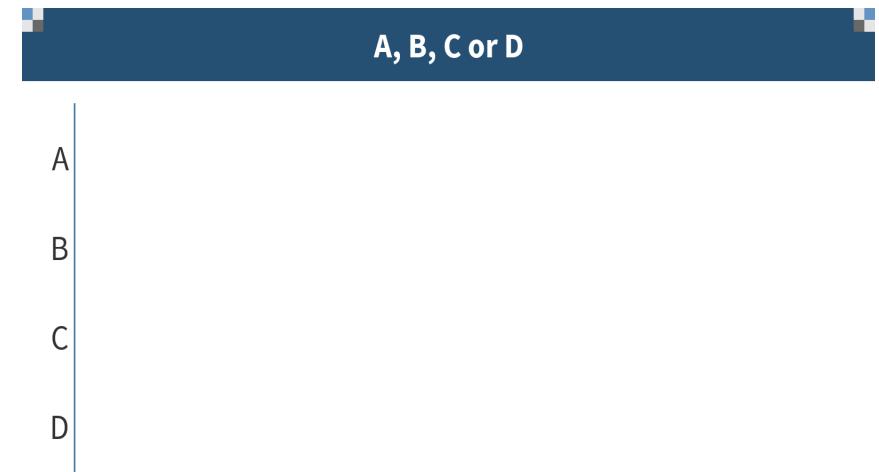
Toy examples

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

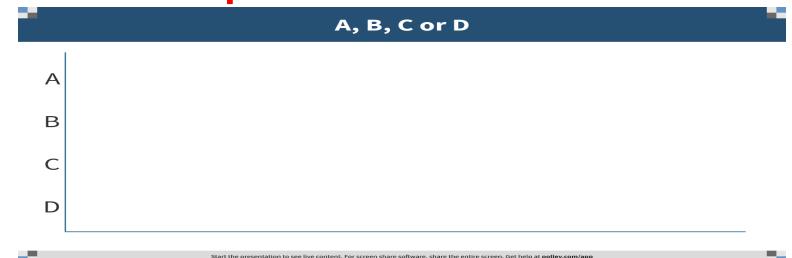
Assume you are learning $y = f(x)$ for x a scalar

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 $-1, -\frac{1}{3}, \frac{1}{3}, 1$
- B) Pick $-1, 1$, see what the answer is, then pick next x
- C) Pick $-\frac{1}{3}, \frac{1}{3}$, 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

◆ Response surface modeling

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

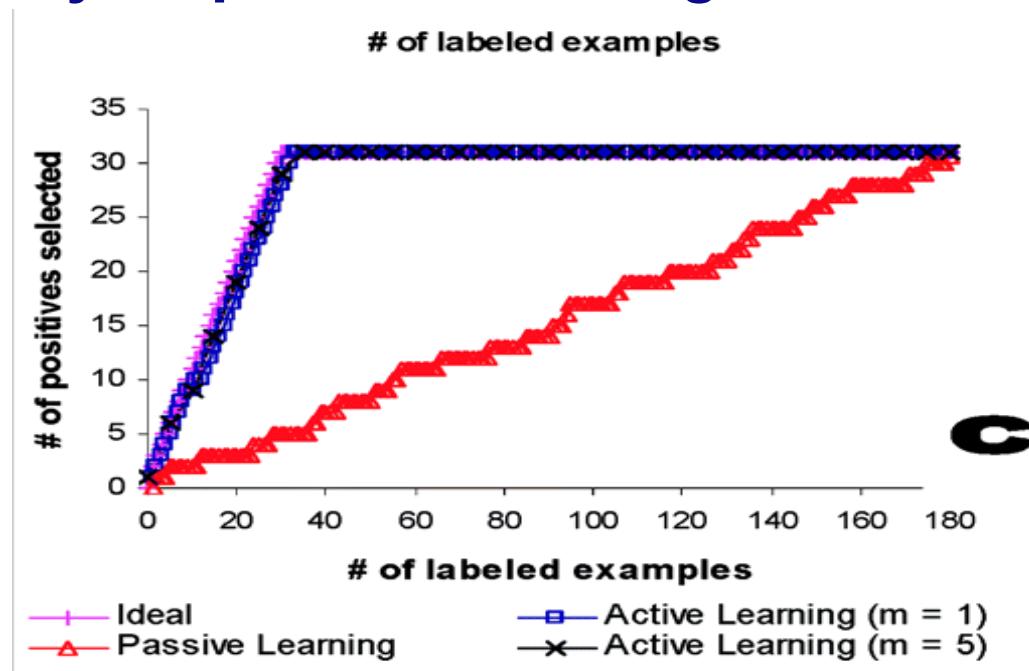
◆ Method:

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

Liu 2004

Example: Gene expression and Cancer classification

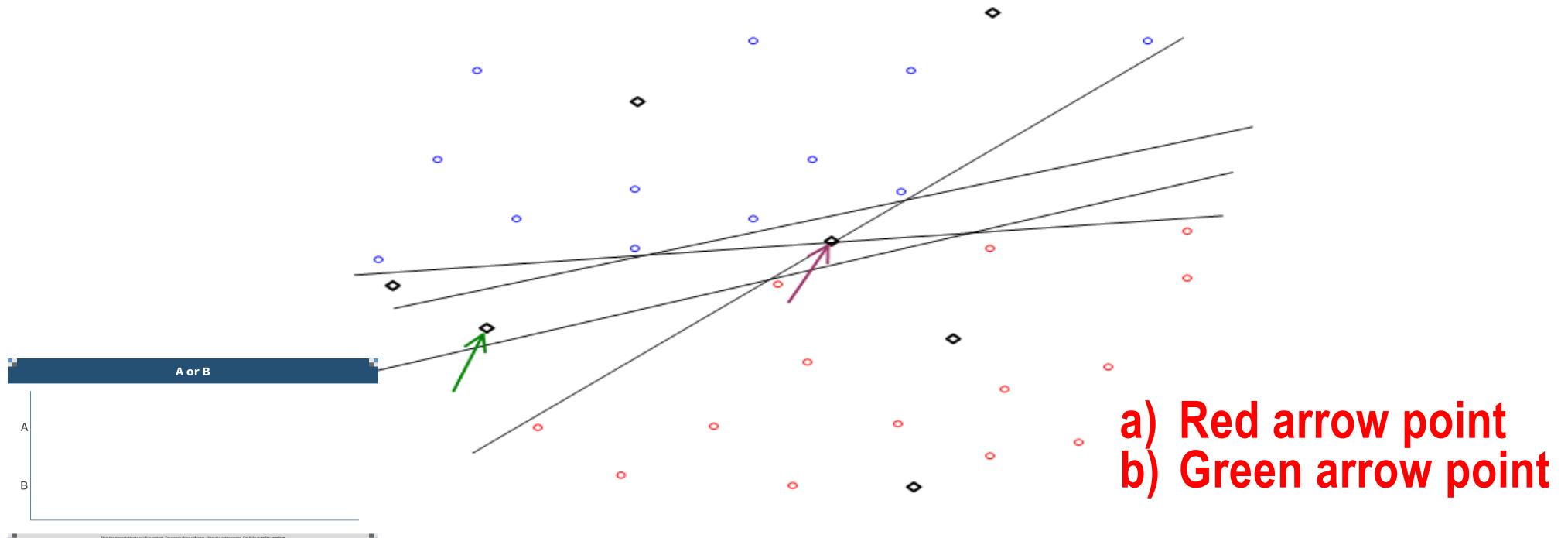
- ◆ Active learning takes 31 points to achieve same accuracy as passive learning with 174



Liu 2004

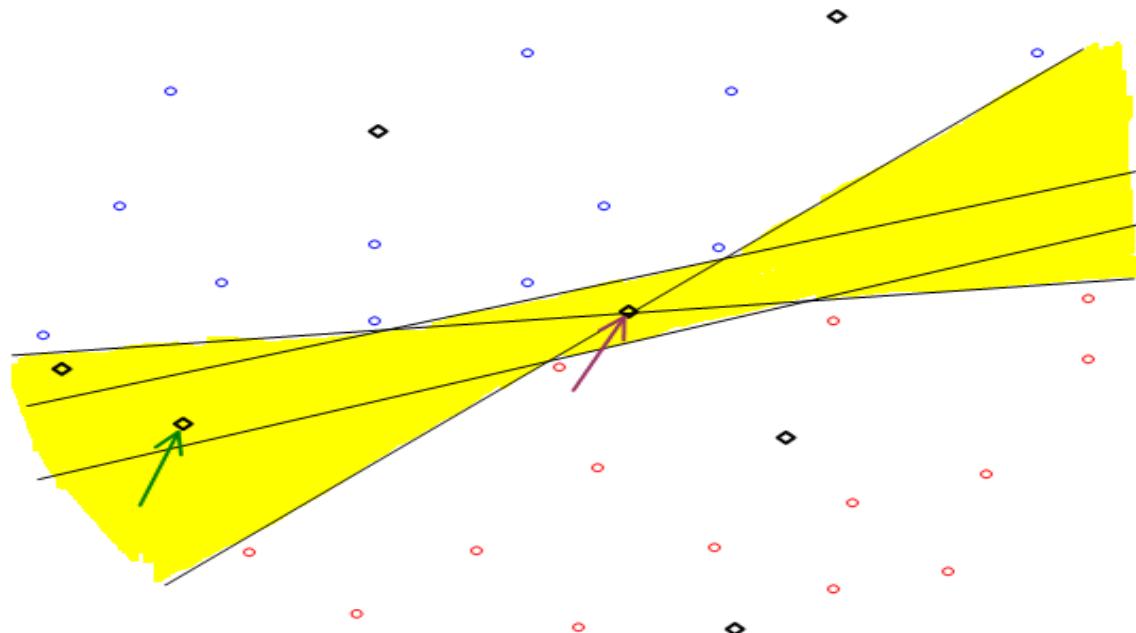
Query by Committee

◆ Which unlabeled point should you choose?



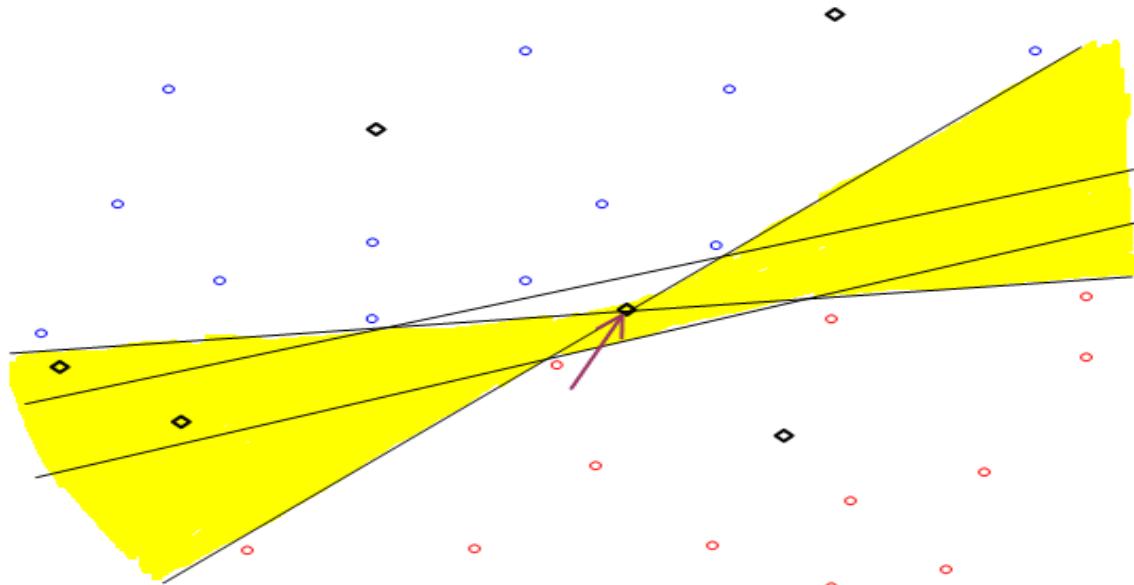
Query by Committee

- ◆ Yellow = valid hypotheses



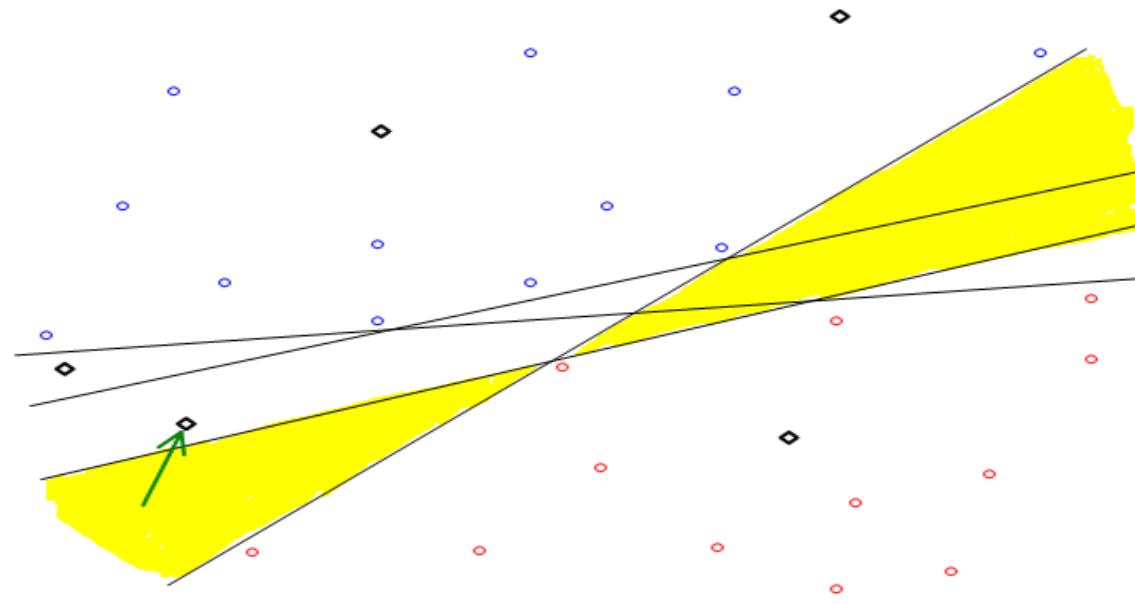
Query by Committee

- ◆ 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

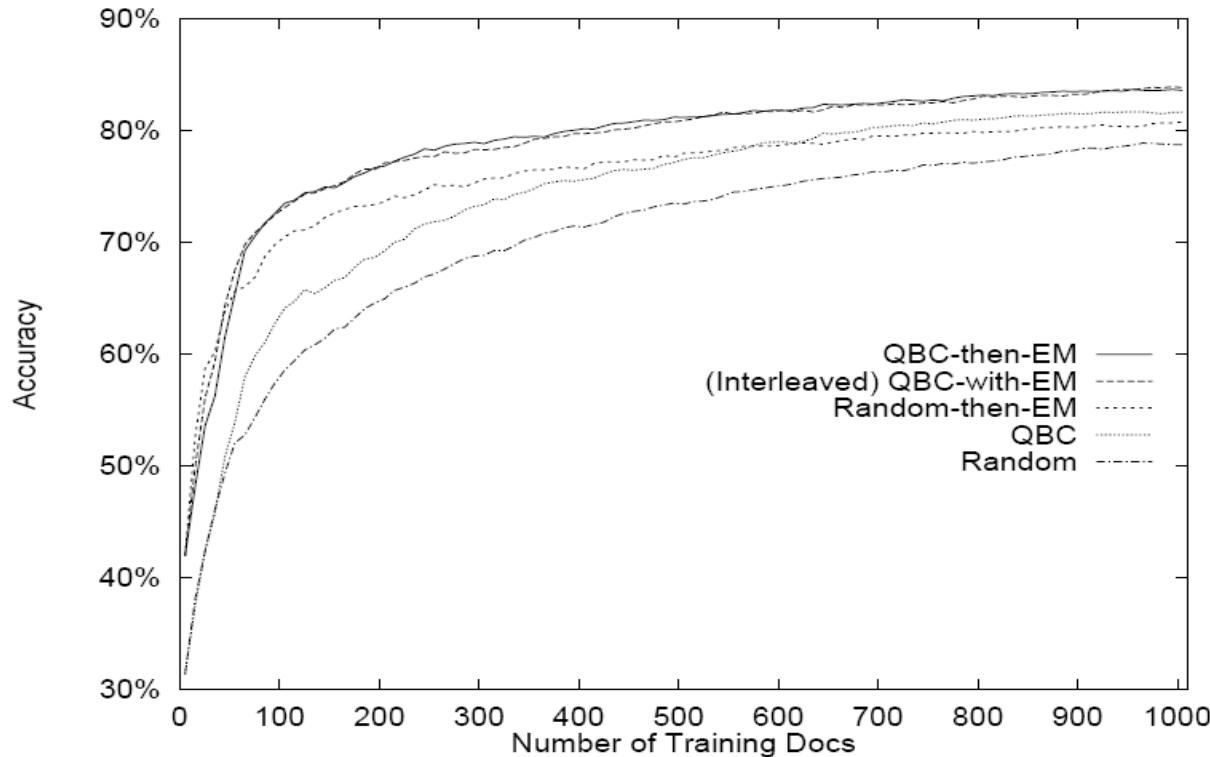


Query by Committee

- ◆ 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
- ◆ Better: quantify the 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)
- ◆ Where have we seen something similar?

[MacKay, 1992]

KL divergence as info gain

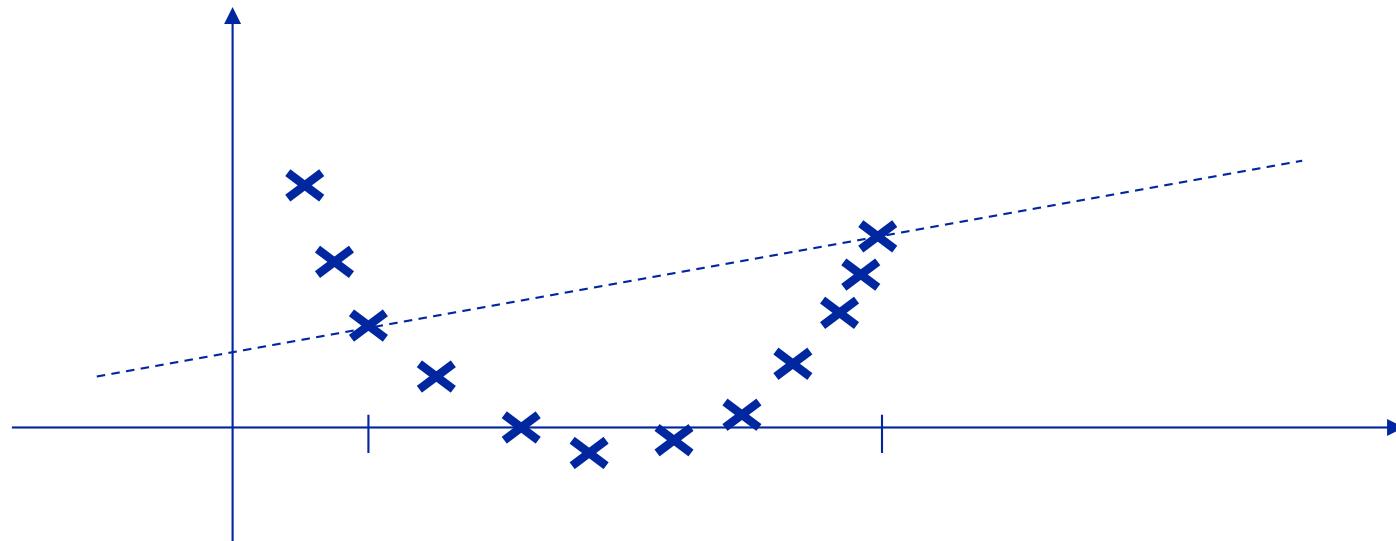
- ◆ The KL divergence measures the information gain expected from query (x'):

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

- ◆ 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**
- ◆ Response surface modeling

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 $y = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon$ then $\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

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

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

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

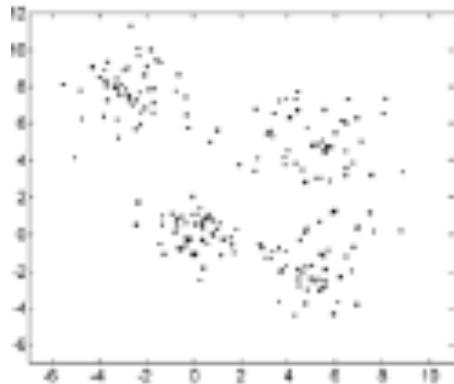
<i>A-optimal</i> (average) design minimizes	$\text{trace}(\mathbf{X}^T \mathbf{X})^{-1}$
<i>D-optimal</i> (determinant) design minimizes	$\log \det(\mathbf{X}^T \mathbf{X})^{-1}$
<i>E-optimal</i> (extreme) design minimizes	max eigenvalue of $(\mathbf{X}^T \mathbf{X})^{-1}$

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

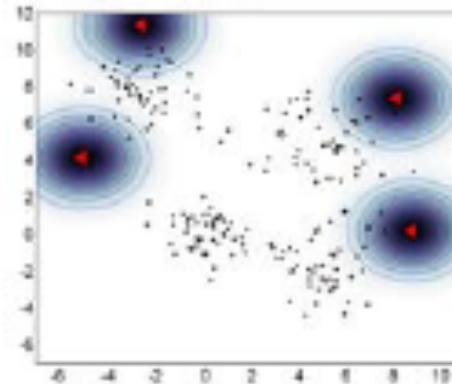
A-Optimal Design

- ◆ ***A-optimal* design minimizes the trace of $(X^T X)^{-1}$**
 - Equivalent to minimizing the Frobenious norm
 - 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]

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
- ◆ **Response surface modeling**

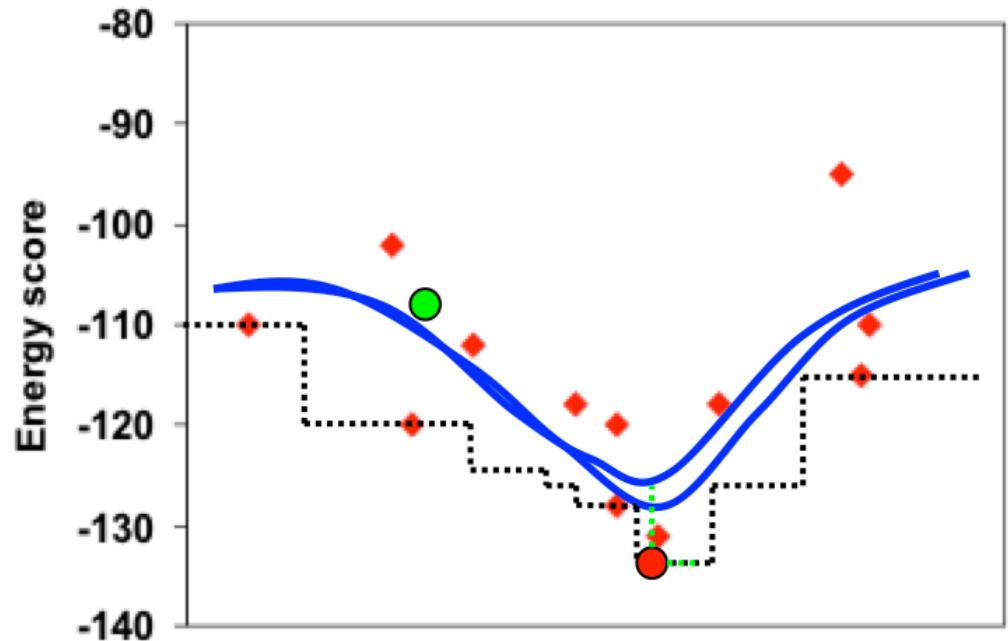
Response Surface Methods

- ◆ Goal: find an \mathbf{x} to maximize y (for an unknown function)
 - Using gradient descent.
 - E.g., find optimal conditions for growing cell cultures
- ◆ Procedure
 - Given a set of datapoints, fit $y = f(\mathbf{x})$ (the “response surface”)
 - E.g., fit a quadratic
 - Hill-climb on the response surface to find the next \mathbf{x}_i
 - Use next \mathbf{x} and y to update the response surface
- ◆ We only care about fitting close to the optimum

Response Surface Modeling

Goal: Approximate the function
 $f(c) = \text{score}(\text{minimize}(c))$

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]

What you should know

- **Active learning (sequential)**
 - Query by committee
 - Uncertainty sampling
 - Information-based loss functions

Sample where models disagree
Sample where model is uncertain
Maximize information gain
- **Optimal experimental design**
 - A-optimal design
 - D-optimal design
 - E-optimal design

Minimize trace of $(X^T X)^{-1}$
Minimize det of $(X^T X)^{-1}$
Minimize largest eigenvalue of $(X^T X)^{-1}$
- **Response surface methods**

Sequential experiments for optimization