Active Learning and Optimized Information Gathering

Lecture 11 – Bayesian Experimental Design

CS 101.2

Andreas Krause

Announcements

- **Homework 2: Due Thursday Feb 19**
- **Project milestone due: Feb 24**
  - 4 Pages, NIPS format: http://nips.cc/PaperInformation/StyleFiles
  - Should contain preliminary results (model, experiments, proofs, …) as well as timeline for remaining work
  - Come to office hours to discuss projects!
- Office hours
  - Come to office hours before your presentation!
  - Andreas: **Monday 3pm-4:30pm**, 260 Jorgensen
  - Ryan: Wednesday 4:00-6:00pm, 109 Moore

Review of Active Learning

- PAC Learning:
  - How many labeled examples do we need to get error $\leq \varepsilon$ with probability $1-\delta$
- Passive learning
  - $n = O'(1/\varepsilon^2(VC(H) + \log 1/\delta))$ suffice
  - Bounds crucially depend on **i.i.d. data**
- Active learning
  - Uncertainty sampling ➔ Bias
  - Can avoid bias (and get fall-back guarantee) using **pool-based** active learning

Algorithms for active learning

- Generalized binary search: Shrink **version space** (set of consistent hypotheses) as quickly as possible
- Sample complexity depends both on H and P_X
  - Splitting index
  - Disagreement coefficient
- Can in some cases get **exponential improvements** in rate of error reduction: $(\log 1/\varepsilon)^2$ instead of $1/\varepsilon^2$ ☺

Course outline

1. Online decision making

2. $\boxed{\text{Statistical active learning}}$

3. Combinatorial approaches

Medical diagnosis

- Want to predict medical condition of patient given noisy symptoms / tests
  - Body temperature
  - Rash on skin
  - Cough
  - Increased antibodies in blood
  - Abnormal MRI

	healthy	*sick*
Treatment	-$$	$
No treatment	0	-$$$

- Treating a healthy patient is bad,
  not treating a sick patient is terrible
- Each test has a (potentially different) cost
- **Which tests should we perform to make most effective decisions?**

General approach:

1. Model patients condition **Y** and outcomes of tests $X_1, \ldots, X_n$ as **random variables**
2. Assign cost for
   - "misdiagnosis" (predicting wrong value of Y)
   - Performing tests (learning value x_i of some X_i)
3. Select tests to perform to minimize total cost

Let's see how we can do this…

Decision theory

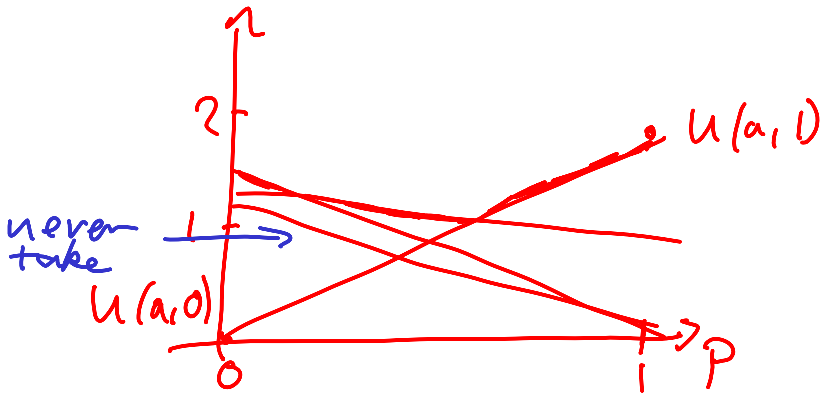
- Bernoulli random variable Y; Y=1 (sick) Y=0 (healthy)
- Can perform two actions: A=1 (treat) or 0 (not treat)
- Obtain utility U(a,y)
- **Don't know y!**

U(a,y)	*Y=0*	*Y=1*
A=0	0	-100
A=1	-10	10

- A priori probability P(Y=1) = p
- Choose action to maximize *expected* utility

$a^* = \text{argmax}_a \, EU(a) =$ $p\,U(a,1) + (1-p)\,U(a,0)$

Shape of expected utility



$p = P(y=1)$

$a: \; p\,U(1,a) + (1-p)\,U(0,a)$

Informed decision making

- Observations help us make decisions
- Model possible observations as random variables $X_1, \ldots, X_n$
- Observing $X_i = x_i$ allows us to perform inference:

$$P(Y=1 \mid X_i = x_i) = \frac{P(Y=1) \cdot P(X_i = x_i)}{P(X_i)}$$

- Observation changes our expected utility (and action)

$$a^* = \text{argmax}_a \; EU(a \mid x_i) = p' \, U(1,a) + (1-p') \, U(0,a)$$

Informed decision making

- More generally, make multiple observations
  $X_1 = x_1, X_4 = x_4, \ldots, X_6 = x_6$
  For index set $B = \{i_1,\ldots,i_k\}$ write $\mathbf{X_B} = (X_{i_1},\ldots, X_{i_k})$

- Compute $P(Y = 1 \mid \mathbf{X_B} = \mathbf{x_B}) = p''$

  ➔ $a^* = \text{argmax}_a \, EU(a \mid \mathbf{x_B}) = p'' \, U(1,a) + (1-p'') \, U(0,a)$

- Value of observing $\mathbf{X_B} = \mathbf{x_B}$:

$$\underbrace{\max_a EU(a \mid \mathbf{x_B})}_{\text{max posterior utility}} - \underbrace{\max_{a'} EU(a')}_{\text{max prior utility}}$$

Value of information [Howard '66]

- Value of observing $\mathbf{X_B} = \mathbf{x_B}$:

$$\text{Value}(X_B = x_B) = \max_a EU(a \mid \mathbf{x_B}) - \max_{a'} EU(a')$$

- But when selecting medical tests $\mathbf{X_B}$ to perform, we don't know their outcome $\mathbf{x_B}$!!

- Bayesian's response:
  Prior belief about likelihood of test outcomes $P(\mathbf{x_B})$

→ Expected value of observing X_B

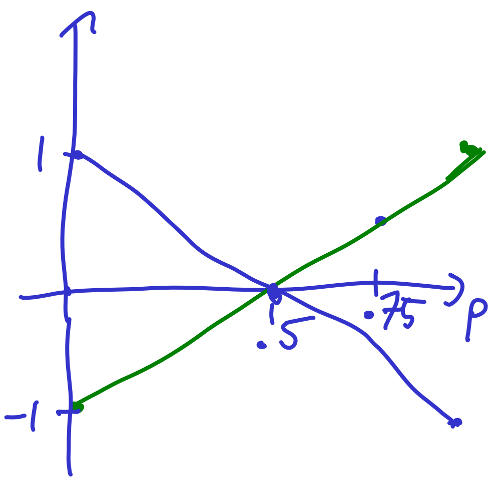
$$\text{VOI}(B) = \sum_{x_B} P(x_B)\, \text{Value}(X_B = x_B)$$

Example value of information

$P(Y) = .5$, $X = Y$ with prob $.5$

$U(\{0,1\})$ with prob $.5$

$$P(Y=1 \mid X=1) = \frac{P(Y=1)\, P(X=1 \mid Y=1)}{P(X=1)} = \frac{\cancel{.5} \cdot \frac{3}{4}}{\cancel{.5}} = \frac{3}{4}$$



$p = P(Y=1)$

	$A=1$	0
$Y=1$	1	-1
0	-1	1

$EU(A=a) = 0$

$EU(A=1 \mid X=1) = .5$

$EU(A=0 \mid X=1) = -.5$

$EU(A=1 \mid X=0) = -.5$

$EU(A=0 \mid X=0) = .5$

$$VOI(X) = P(X=1) \cdot \max_a EU(A=a \mid X=1) + P(X=0) \; \ldots$$

$= .5 \cdot .5 + .5 \cdot .5 = .5$

Greedy Information gathering

- Start with no observations B={};
- $V = \max_a EU(a)$
- Repeat
  - For each test X_i compute

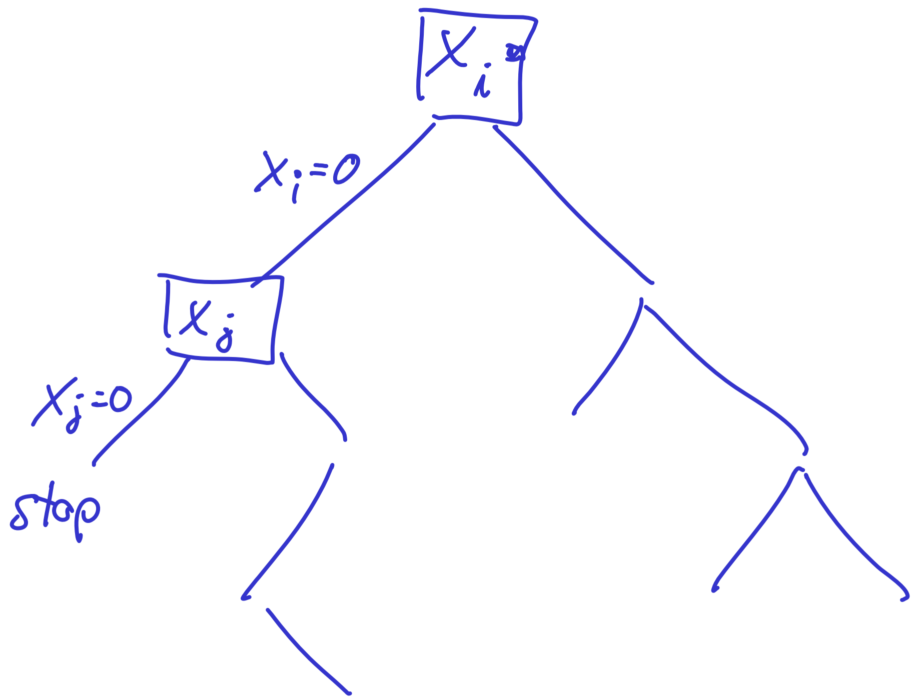
    $p_i = P(X_i = 1 \mid \mathbf{X}_B = \mathbf{x}_B)$

    $V_i = p_i \max_a EU(a \mid X_i{=}1, \mathbf{x_B}) + (1\text{-}p_i) \max_a EU(a \mid X_i{=}0, \mathbf{x_B})$

  - Let $i^* = \text{argmax}_i\, V_i$
  - If $V_{i*} \leq V$ then break

    Else observe $X_{i*} = x_{i*}$; $B = B \cup \{i^*\}$

Decision trees



Greedy algorithm optimal??

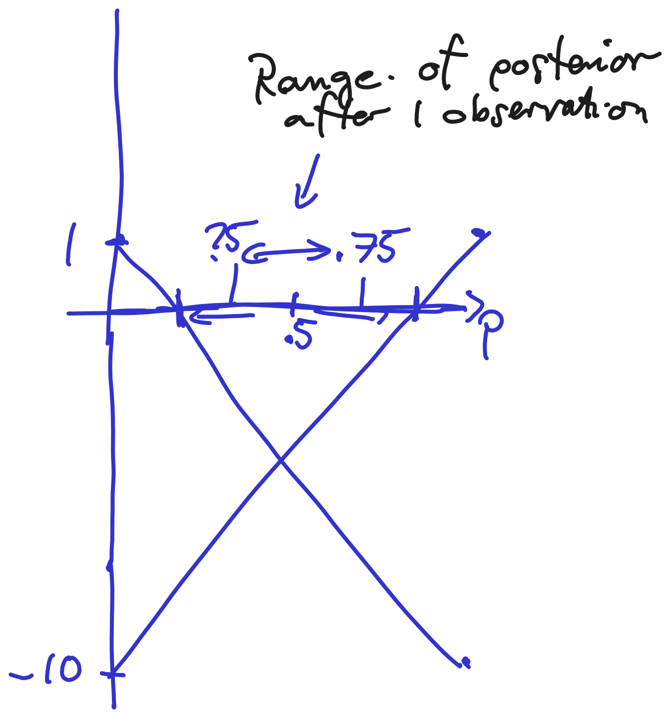
$P(Y=1) = .5$

$X_1 \ldots X_m$

$X_i = Y$ with prob .5

$U\{0,1\}$ with prob .5 ←

$X_1 \ldots X_m$ conditionally independent given Y



U	A=1	A=0	A=pass
Y=1	1	−10	$-\varepsilon$
Y=0	−10	1	$-\varepsilon$

$P(Y=1 \mid X_1=1, X_2=1, \ldots X_6=1) = 1 - .5^{26}$

Optimal value of information

- Can we **efficiently** find an optimal decision tree?

➔ Answer depends on properties of the distribution $P(X_1,\ldots,X_n,Y)$

$P(X_1 \ldots X_n) = \prod_i P(X_i \mid X_{i-1})$

Theorem [Krause & Guestrin IJCAI '05]:

- If the random variables form a Markov Chain, can find optimal (exponentially large!) decision tree in polynomial time ☺

- There exists a class of distributions for which we can perform efficient inference (i.e., compute $P(Y|X_i)$), where finding the optimal decision tree is **NP^{PP} hard**

Approximating value of information?

- If we can't find an optimal solution, can we find **provably near-optimal** approximations??

- Yes, but have to make certain assumptions about the value of information objective (next 2 lectures)

Generalizing value of information

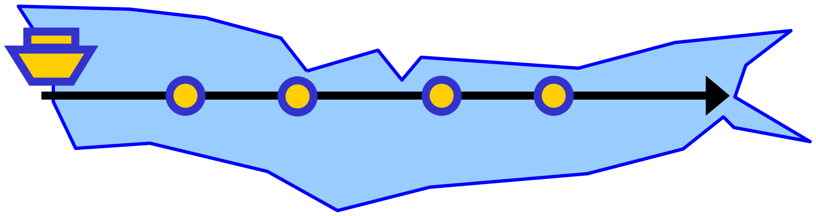
Prior P(Y) **obs $X_i = x_i$** **Posterior P(Y | x_i)** **Reward**

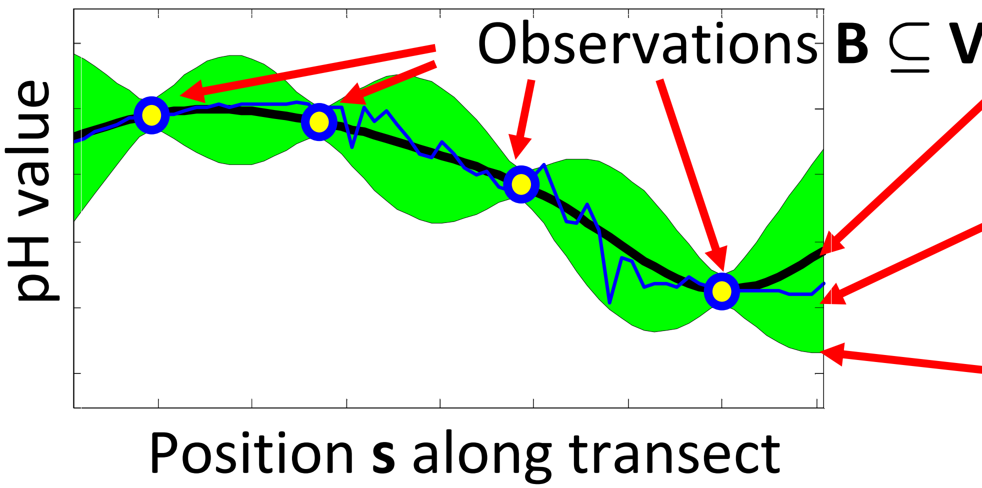
- Value of information:
  Reward[P(Y | x_i)] = $\max_a$ EU(a | x_i)

- Reward can by **any function** of the distribution P(Y | x_i)
- Important examples:
  - Posterior variance of Y
  - Posterior entropy of Y

Automated environmental monitoring

- Monitor pH values using robotic sensor transect





Prediction at unobserved locations

True (hidden) pH values

Use **probabilistic model** (Gaussian processes) to estimate prediction error

Recap: Gaussian processes

- A Gaussian Process (GP) is a
  - (infinite) set of random variables, indexed by some set V
    i.e., for each $x \in V$ there's a RV Y_x
  - Let $A \subseteq V$, $|A| = \{x_1, \ldots, x_k\} < \infty$
    Then

    $$Y_A \sim N(\mu_A, \Sigma_{AA})$$

    where

$$\Sigma_{AA} = \begin{pmatrix} \mathcal{K}(x_1,x_1) & \mathcal{K}(x_1,x_2) & \ldots & \mathcal{K}(x_1,x_n) \\ \vdots & & & \vdots \\ \mathcal{K}(x_k,x_1) & \mathcal{K}(x_k,x_2) & \ldots & \mathcal{K}(x_k,x_k) \end{pmatrix} \qquad \mu_A = \begin{pmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_k) \end{pmatrix}$$

  - K: $V \times V \to R$ is called **kernel** (covariance) function
    μ: $V \to R$ is called **mean** function

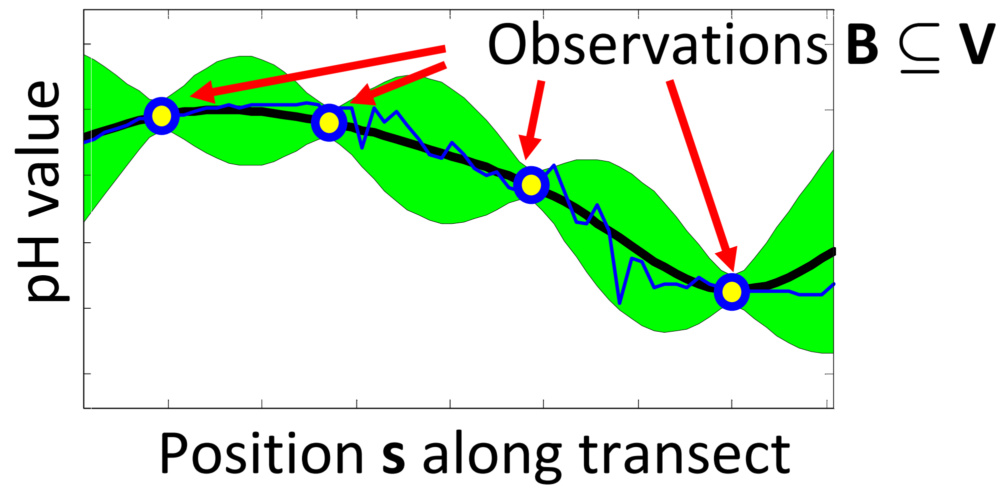
Inference in GPs

- Set of locations V
- Observations $X_B = x_B$ at locations B
- Want to make predictions at unobserved locations A
- $P(X_A = x_A | X_B = x_B) = N(x_A; \mu_{A|B}, \Sigma_{A|B})$

$$\mu_{A|B} = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B)$$

$$\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$$

Spatial prediction in GPs



World discretized into finite set of locations **V**

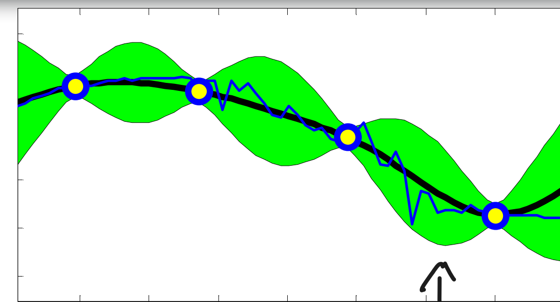
- Based on observations $X_B = x_B$ at locations B, make predictions at unobserved locations $A \subseteq V$:

  $P(X_A = x_A \mid X_B = x_B) = N(x_A; \mu_{A|B}, \Sigma_{A|B})$

In order to select most useful observations, need to **quantify uncertainty** in predictive distribution $P(X_A \mid x_B)$

Quantifying uncertainty

- Different possibilities used in practice:



"Volume" of confidence band

- Expected mean squared prediction error (EMSE):

  $$\mathrm{EMSE}(X_A \mid X_B = x_B) = 1/|B| \sum_s \sigma_{s|A}^2$$ ← Average posterior variance

- Maximum predictive variance (MPV):

  $$\mathrm{MPV}(X_A \mid X_B = x_B) = \max_s \sigma_{s|A}^2$$

- Entropy: $H(X_A \mid x_B) = -\int p(x_A \mid x_B) \log p(x_A \mid x_B)\, dx_A$

  $$H(X_A \mid X_B = x_B) = \tfrac{1}{2} \underline{\log |\Sigma_{A|B}|} + n/2 \log (2 \pi e)$$

Greedy Bayesian experimental design

- Start with no observations B={};
- For i = 1 to k
  - For each possible location X_i compute

    V_i = EMSE(X_A | X_i, X_B)

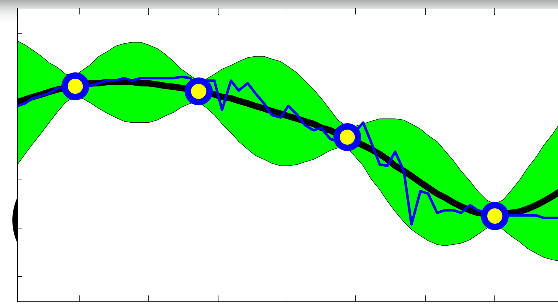
  - Let i* = $\text{argmin}_i V_i$
  - Observe $X_{i*} = x_{i*}$; $B = B \cup \{i^*\}$

Greedy algorithm

- Matlab demo

Quantifying uncertainty

- Different possibilities:



- Expected mean squared prediction error (
  “Bayesian A-optimality”

  $$\text{EMSE}(X_A \mid X_B = x_B) = 1/|B| \sum_s \sigma_{s|B}^2$$

$$\Sigma_{B|A} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1m}^2 \\ \vdots & & \ddots & \vdots \\ \sigma_{m1}^2 & \cdots & & \sigma_{mm}^2 \end{pmatrix}$$

- Maximum predictive variance (MPV):

  $$\text{MPV}(X_A \mid X_B = x_B) = \max_s \sigma_{s|A}^2$$

EMSE $= \operatorname{trace} \Sigma_{B|A}$

does not depend on $\mu_{B|A}$

- Entropy: **“Bayesian D-optimality”**

  $$H(X_A \mid X_B = x_B) = ½ \log |\Sigma_{A|B}| + n/2 \log (2 \pi e)$$

All these measures do ONLY depend on $\Sigma_{B|A}$

Independence of observations

- EMSE, MPV, Entropy only depend on $\Sigma_{A \mid B}$

$$\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$$

does not depend on actual observations $X_B = x_B$

- $F_{EMSE}(B) = \int p(x_B) \underbrace{EMSE(X_A \mid X_B = x_B)}_{\text{does not depend on } x_B} dx_B$

  $= 1/n \text{ trace } \Sigma_{A|B}$ trace $\Sigma_{A|B}$

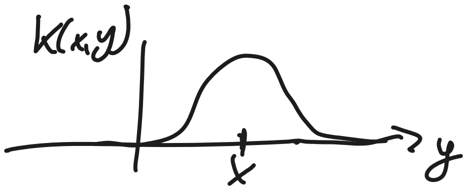
- Expected reward when observing B independent of actual observation x_B!

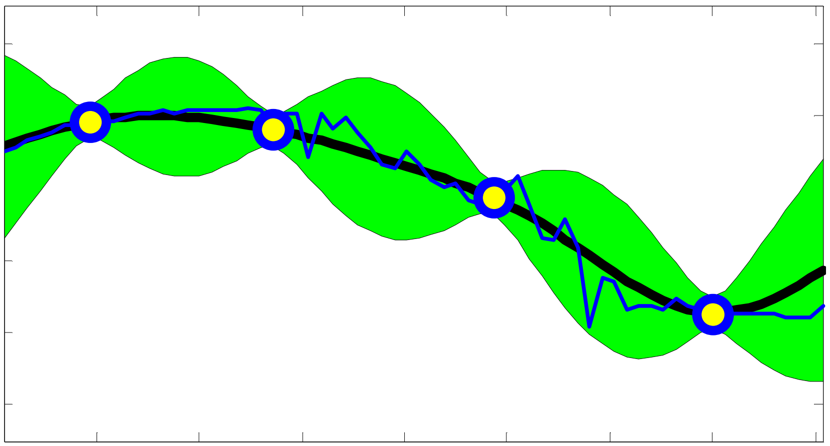
- Expected posterior EMSE only depends on chosen locations B!

Implications

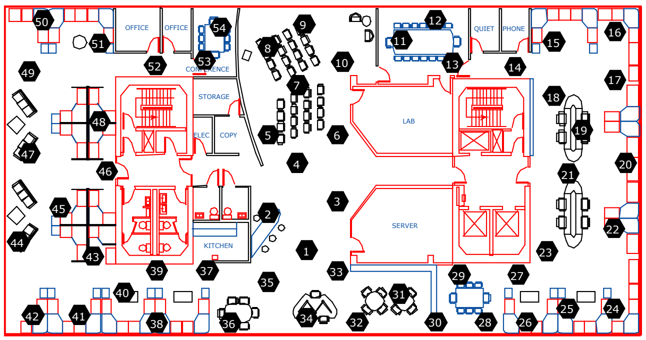
- Can plan observations ahead of time before making measurements (logistically simpler)

- If kernel is isotropic $K(x,y) = f(\ |x-y|\ )$, regularly spaced designs are optimal
  Example: $K(x,y) = \exp(\ (x-y)^2/\theta^2)$

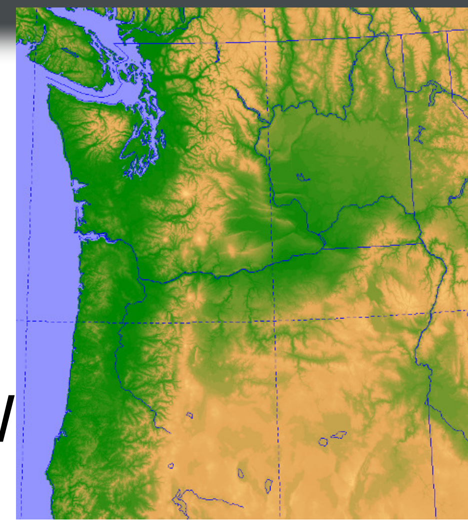


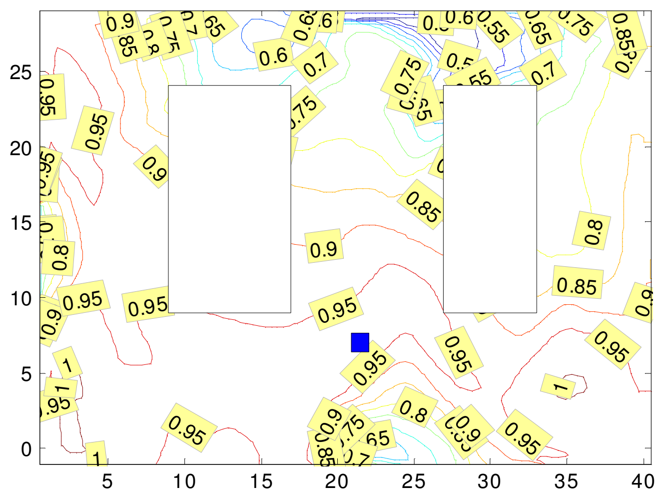


Nonstationary spatial correlation

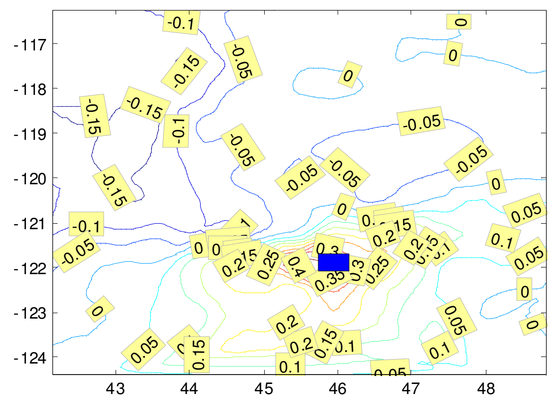


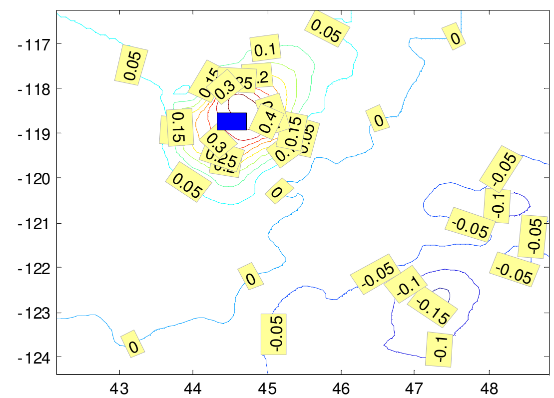
Precipitation (rain) data from Pacific NW





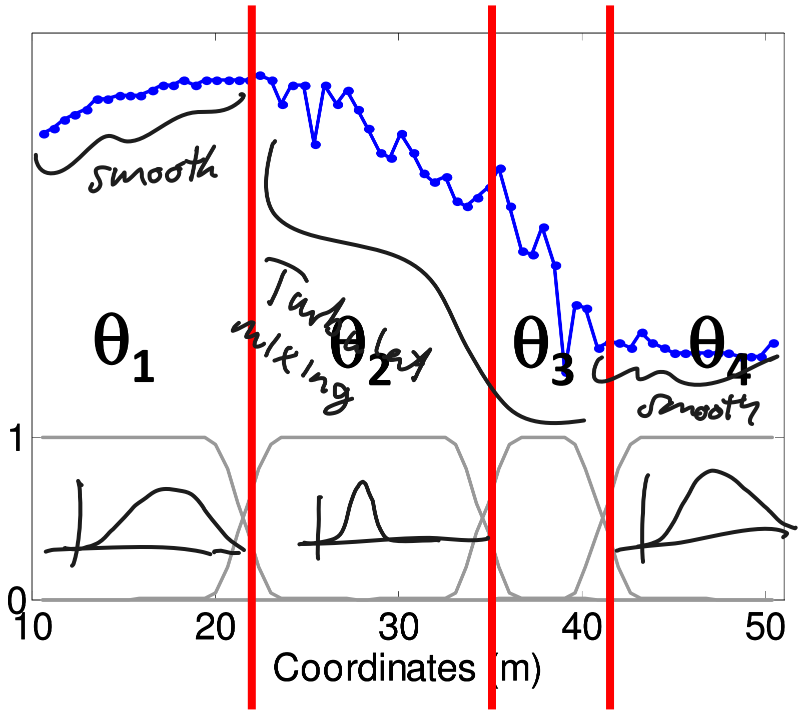
Non-local, Non-circular correlations





Complex positive and negative correlations

Nonstationarity by spatial partitioning

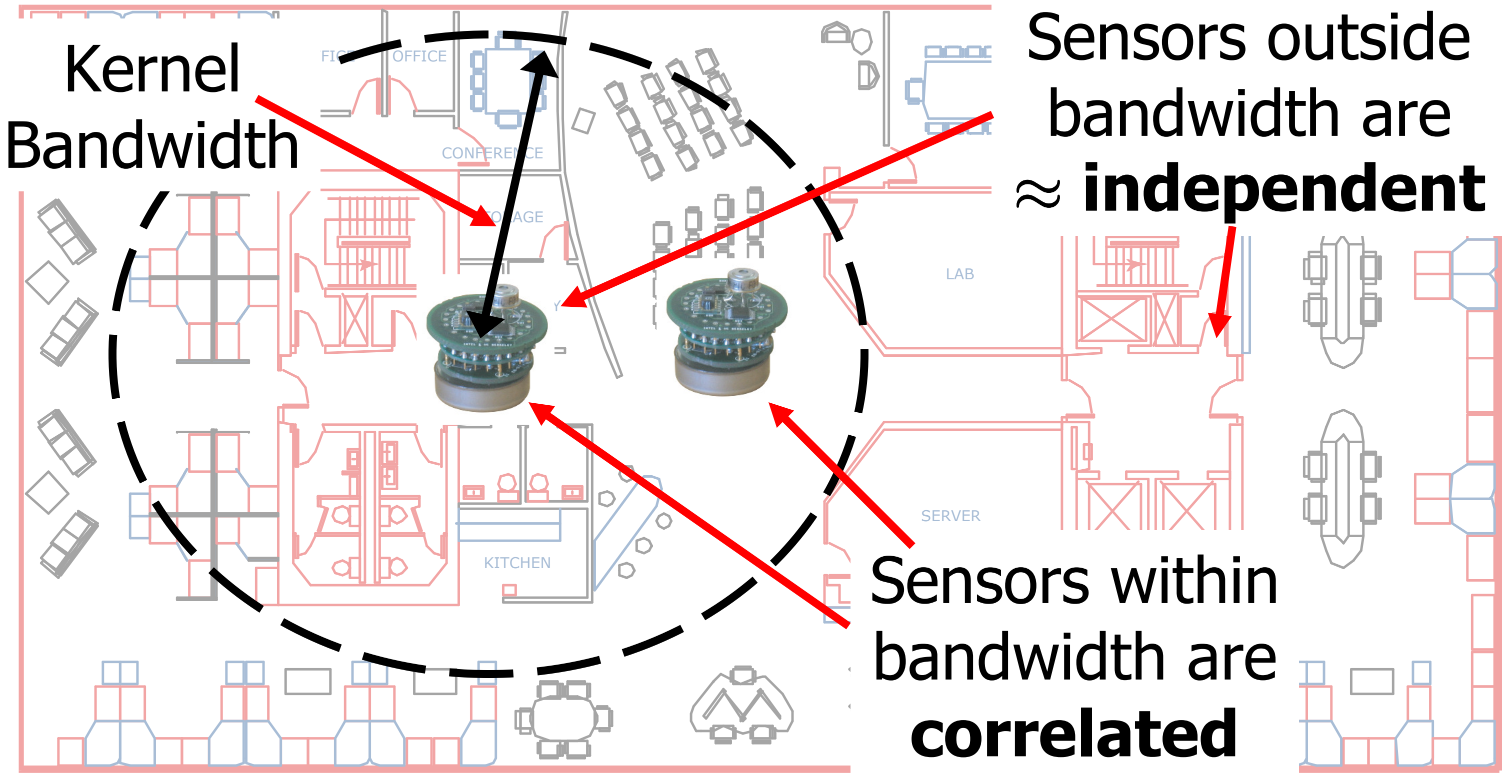


- **Partition** into regions
- **Isotropic** GP for each region, **weighted** by region membership

  $K(x,y) = \exp((x-y)^2/\theta_i^2)$
- Final GP is spatially varying **linear combination**

- Need to learn parameters θ_i of nonstationary kernel function from data
- Can apply techniques from active learning to do that [Krause & Guestrin, ICML '07]

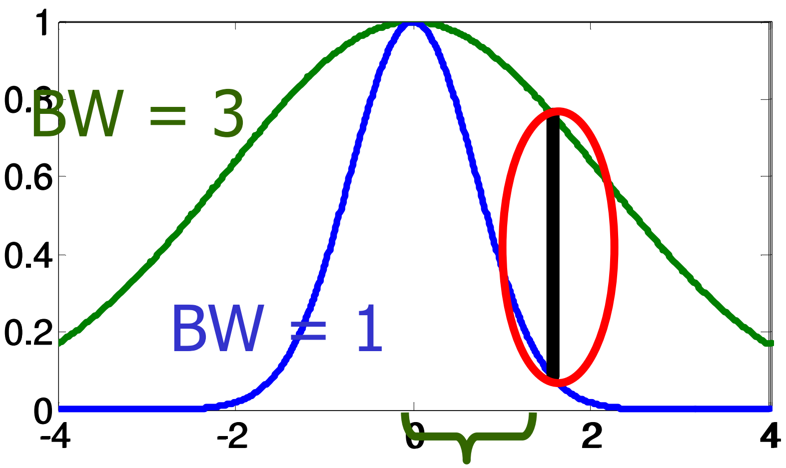
Learning the bandwidth



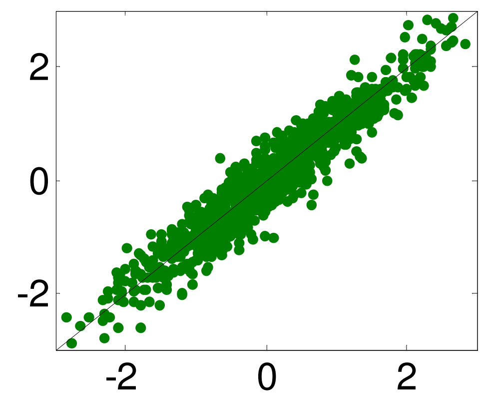
Can **narrow down** kernel **bandwidth** by sensing **within** and **outside** bandwidth distance! ☺

Hypothesis testing: Distinguishing two bandwidths

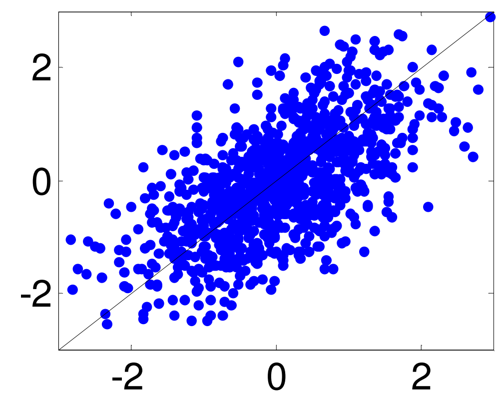
- Square exponential kernel: $\mathcal{K}(s,t) = \exp\left(-\frac{\|s-t\|_2^2}{\theta^2}\right)$



Distance Δ
correlation gap largest



Correlation under BW=3

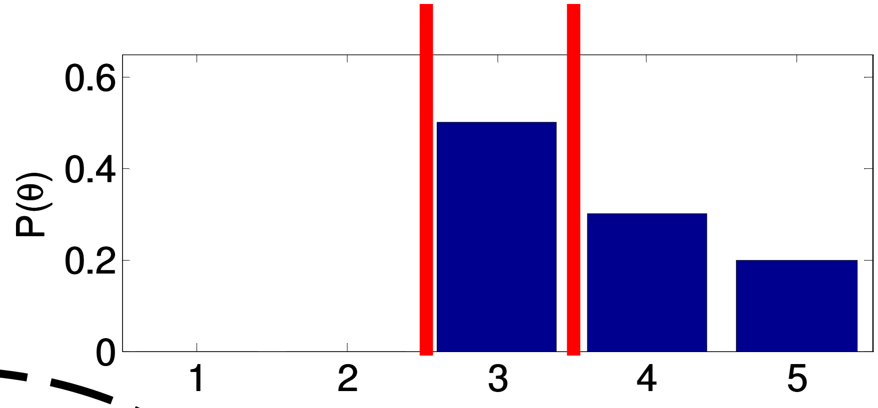


Correlation under BW=1

- Choose pairs of samples at distance Δ to **test correlation!**

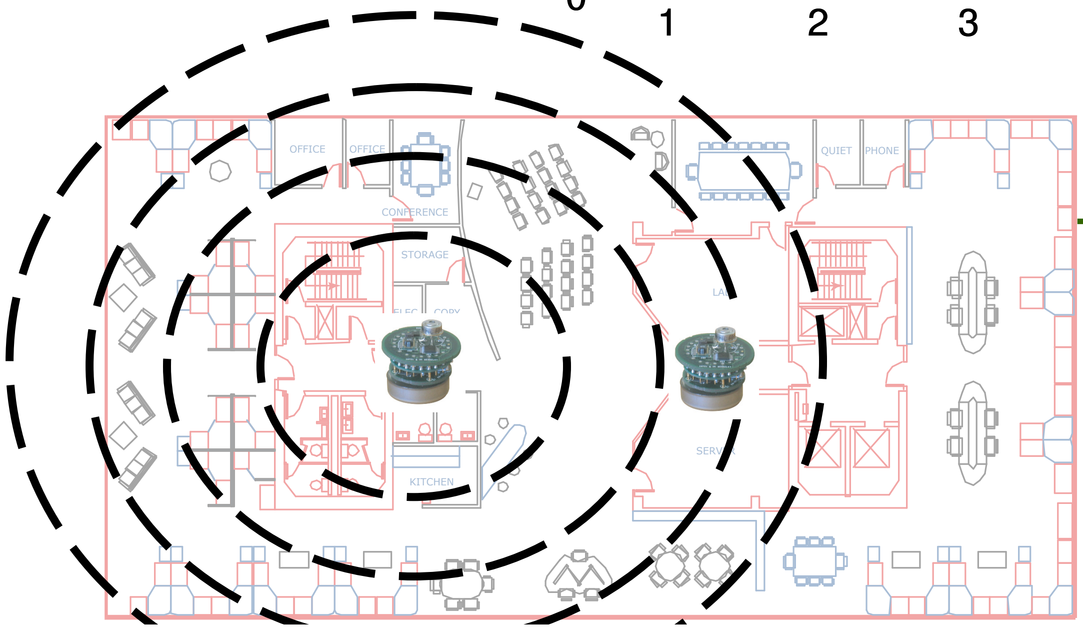
Hypothesis testing: Searching for bandwidth

- Find "**most informative split**" at posterior median



Test: BW>2?

Test: BW>3?



Testing **policy** needs only **logarithmically** many tests! ☺

What you need to know

- Maximum expected utility principle
- Value of information
- Bayesian experimental design in GPs
  - Bayesian active learning for regression
  - Different optimality criteria (EMSE, MPV, Entropy)