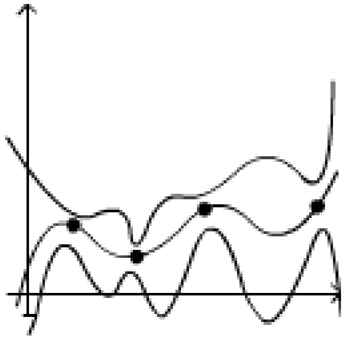
CS/CNS/EE 253: Advanced Topics in Machine Learning
Topic: Uncertainty Sampling **Lecturer:** Andreas Krause
Scribe: Daniel Rosenberg **Date:** March 8, 2010

17.1 Uncertainty Sampling



When choosing a point to sample when attempting to learn a function, it is important to choose the 'most informative point'. One interpretation of this is to choose the point that currently has the highest variance:

$$x_{t+1} \in \operatorname*{argmax}_{x \in X} \sigma_t^2(x)$$

The utility function $F : 2^X \to \mathbb{R}$ for $A \subset X, |A| = k$ for this method can be written in terms of mutual information, or in terms of entopy as follows:

$$\begin{aligned} F(A) = I(f; y_A) &= H(y_A) - H(y_A|f) \\ &= H(f) - H(f|y_n) \end{aligned}$$

For a gaussian process, the marginal entropy is given by:

$$H(y_A) = \lg(2\pi e)^{-\frac{k}{2}} \sqrt{|\Sigma_{AA} + \sigma^2 I|}$$

and the conditional entropy is given by:

$$H(y_A|f) = \log(2\pi e)^{-\frac{k}{2}} \sqrt{|\sigma^2 I|}$$

17.1.1 Properties of H

Let x, y, z be random variables in $\mathbb{R}$

1. $H(x|y) \geq H(x|y,z)$ "Information never hurts"
2. $H(x,y) = H(x) + \underbrace{H(y|x)}_{\int p(x=x')H(y|x=x')\mathrm{d}x'}$ "Chain rule"

Note that the entropy can be negative. Consider the 1D Gaussian. The entropy is given by $H(y) = \log\sqrt{2\pi e\sigma^2}$, and log will be negative if σ is small enough.
Given a set A, we wish to know what point we should choose to maximize our gain in utility.

$$A = \{x_1 \dots x_t\}$$

$$\begin{aligned} F(A \cup \{x\}) - F(A) &= H(y_A, x) - H(y_{A,x}|f)) \\ &\quad - [H(y_A) - H(y_A|f)] \\ &= H(y_x|y_A) - \text{constant} \\ \operatorname*{argmax}_x H(y_x|y_A) &= \operatorname*{argmax}_x \sigma_t^2(x) \end{aligned}$$

This results in choosing the point with the highest variance, or the most uncertain point.

17.1.2 Analysis of US

Key Question : Uncertainty Sampling constructs $A_k = \{x_1, \dots, x_k\}$.
How does $F(A_k)$ compare to $\max_{|A|\leq k} F(A)$?

We will show : $F(A_k) \geq (1 - \frac{1}{e}) \max_{|A|\leq k} F(A)$

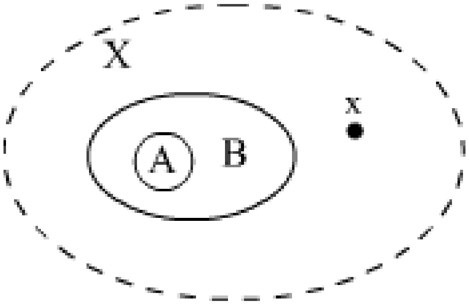


Figure 17.1.1: Relationship between A, B, X, and x

Definition: $G : 2^x \to \mathbb{R}$ is called submodular if $\forall A \subset B \subset X, x \in X \backslash B$
$F(A \cup \{x\}) - F(A) \geq F(B \cup \{x\}) - F(B)$

F is monotonic if $\forall A \subset B \subset X; F(A) \subseteq F(B)$

Claim: $F(A) = I(f; y_n)$ is monotonic and submodular.

Proof:

F is monotonic:

Consider $B = A \cup C$

$$F(A) = H(f) - H(f|y_A)$$

$$A \subset B : F(B) = H(f) - \underbrace{H(f|y_B)}_{\leq H(f|y_A)} \geq F(A)$$

$$F(B) \geq F(A) \therefore F \text{is monotonic}$$

F is submodular:

$$F(A \cup \{x\}) - F(A) - (F(B \cup \{x\}) - F(B))$$

$$H(y_x|y_A) - constant - (H(y_x|y_B) - constant)$$

$$H(y_x|y_A) - \underbrace{H(y_x|y_B)}_{\leq H(y_x|y_A)}$$

So, using "Info Never Hurts" we can see that $F(A \cup \{x\}) - F(A) \geq F(B \cup \{x\}) - F(B)$, and therefore F must be submodular as well. ■

Theorem 17.1.1 *[Nemhauser, Fisher, Wolsey '78]*
Suppose $F : 2^X \to \mathbb{R}$ is monotonic and submodular, $f(\emptyset) = 0$. Then for the greedy solution, A_k, it holds that $F(A_k) = (1 - \frac{1}{e}) \max_{|A| \leq k} F(A)$

This means that Uncertainty Sampling is near-optimal with respect to $I(y_A, f)$.

17.2 Issues with US (Maximum entropy sampling)

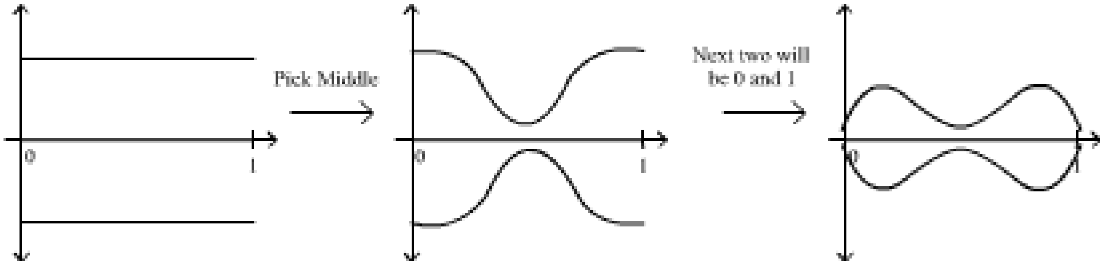


Figure 17.2.2: The first 3 points chosen by MES

When using maximum entropy sampling, you will always choose the boundary values, since those become the most uncertain points. It may be desirable to do something different, for example, choosing equally spaced points. It is worthwhile to consider alternative objective functions.

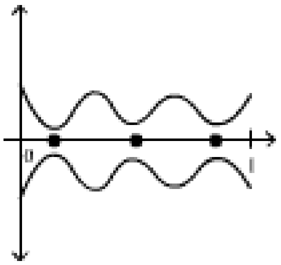


Figure 17.2.3: The confidence band has less area in this case if you select points evenly across the domain.

17.2.1 Bayesian Experimental Design in Gaussian Processes

- US/MES: This is the method we've discussed, where you sample at the most uncertain location. It is also known as D-optimality. $F(A) = H(f) - H(f|y_A) = I(y_A; f)$
- Mutual Information: Pick the point that gives the maximum amount of information. $F_{MI}(A) = I(y_A; y_{X \setminus A})$

  $$F_{MI}(A \cup \{x\}) - F_{MI}(A) = H(y_x|y_A) - H(y_x|y_{\bar{A}})$$

  Here, $y_{\bar{A}} = X \setminus A \setminus \{x\}$
- Variance Reduction: Pick the point that reduces the volume of the confidence band the most. $F_V(A) = \int (\sigma^2(x) - \sigma^2(x|A)) \mathrm{d}x$

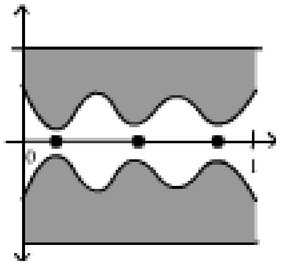


Figure 17.2.4: F_V maximized the difference in volumes of the variance.

- Worst Case Variance Reduction: Picks the point that reduces the max variance the most. $F_W(A) = \max_x \sigma^2(x) - \max_x \sigma^2(x|A)$

F_H, F_{MI}, and F_V (under specific circumstances) are (approximately) monotonic and submodular. F_W is highly nonsubmodular, so the greedy algorithm will perform arbitrarily poorly. There are, however, specialized algorithms that do work well for F_W.

F_H, F_{MI},F_V,F_W, only depend on the predictive covariance, $\Sigma_{x|A}$.

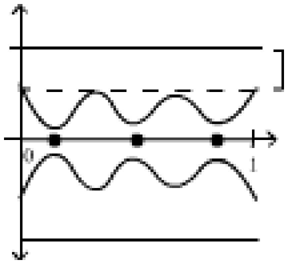


Figure 17.2.5: The indicated distance is the value that F_W maximizes.

$$F_H(A) = H(f) - H(f|y_x) = \log|\Sigma_{xx}| - \log|\Sigma_{x|A}|$$
$$F_V(A) = \text{trace}(\Sigma_{xx}) - \text{trace}(\Sigma_{x|A})$$

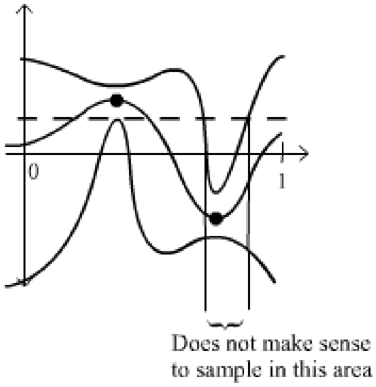
$$\text{For Gaussians, } \Sigma_{x|A} = \Sigma_{xx} - \Sigma_{xA}(\Sigma AA + \sigma^2 I)^{-1}\Sigma_{Ax}$$
$$\mu_{x|A} = \Sigma_{xA}(\Sigma AA + \sigma^2 I)^{-1} y_A$$

$\Sigma_{x|A}$ does NOT depend on y_A

17.3 GP Maximization

Suppose we have a complicated machine with many different knobs. How do we set knobs to make the machine work well? We don't care about the machine's performance in non-optimal regions, we just want it to run as quickly as it can. This can be thought of as attempting to optimize a function f.

Suppose we want to optimize f. We pick $x_1, \ldots, x_t$ such that $f(x_1), \ldots, f(x_t)$ converges to $\max_{x \in X} f(x)$ as quickly as possible. The problem differs from the previous one in that we only wish to find the maximum, and do not care how the function performs in suboptimal areas.



Naive idea: $x_{t+1} \in \operatorname{argmax}_{x \in X} \mu_t(x)$
What's wrong with this idea?
This will get stuck in local maxima, continuously picking the current highest average point.

17.3.1 GP Maximization selection rules

There are many different approaches to this problem. They all attempt explore regions where they are likely to find the function maximum.

$$\text{let } u = \max_{x \in X} \mu_t(x)$$

- Upper Confidence Bound (UCB):

$$x_{t+1} \in \operatorname*{argmax}_{x \in X} \beta_t \sigma_t(x) + \mu_t(x)$$

- Most Probable Improvement (MPI):

$$x_{t+1} = \operatorname*{argmax}_{x \in X} \mathbf{Pr}[f(x) \geq u]$$

- Maximum Expected Improvement (MEI):

$$x_{t+1} = \operatorname*{argmax}_{x \in X} \mathbf{E}[\max f(x) - u, 0]$$

17.4 Next Lecture : Analyzing UCB

As in the bandit case, we will analyze the UCB algorithm by looking at its regret. We can start off by considering the instantaneous regret, given by:

$$r_t = f(x^*) - f(x_t)$$

The total regret is then a summation of these instantaneous regrets.

$$R_T = \sum_{t=1}^{T} r_t$$

We wish to show that $\frac{1}{T} R_T \to 0$ as quickly as possible.

$$f(x^*) - \max\{f(x_1, \ldots, x_t)\} \leq \frac{1}{T} R_T$$

Next lecture, we will prove the following bound on the regret of UCB:

If $|x| = n$, you have a kernel $k : X \times X \to \mathbb{R}$, and $f \sim \mathrm{GP}(0, k)$, then for the regret of UCB, ot holds that:

$$R_T = O^*(\sqrt{T \beta_T \gamma_T})$$

$$\beta_T = O(\log t)$$

where γ_T is worst case experimental gain.

$$\gamma_T = \max_{|A| \leq T} I(f; y_A)$$

For SE kernel $k(x, x') = \exp \frac{(x-x')}{h'^2}$

$$\gamma_T = O^*((\log T)^d))$$

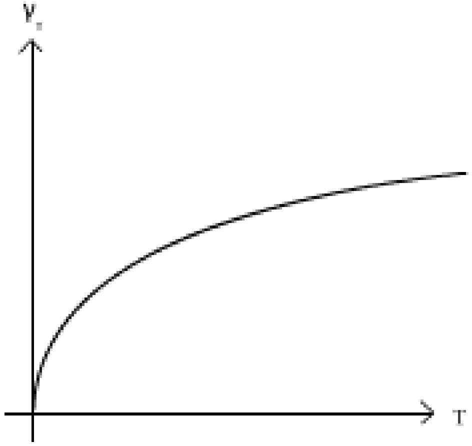


Figure 17.4.6: As a result of submodularity, we experience a diminishing return on γ_T. Although it increases, the rate of increase diminishes.