Treed Gaussian Processes and Adaptive Sampling

Herbie Lee

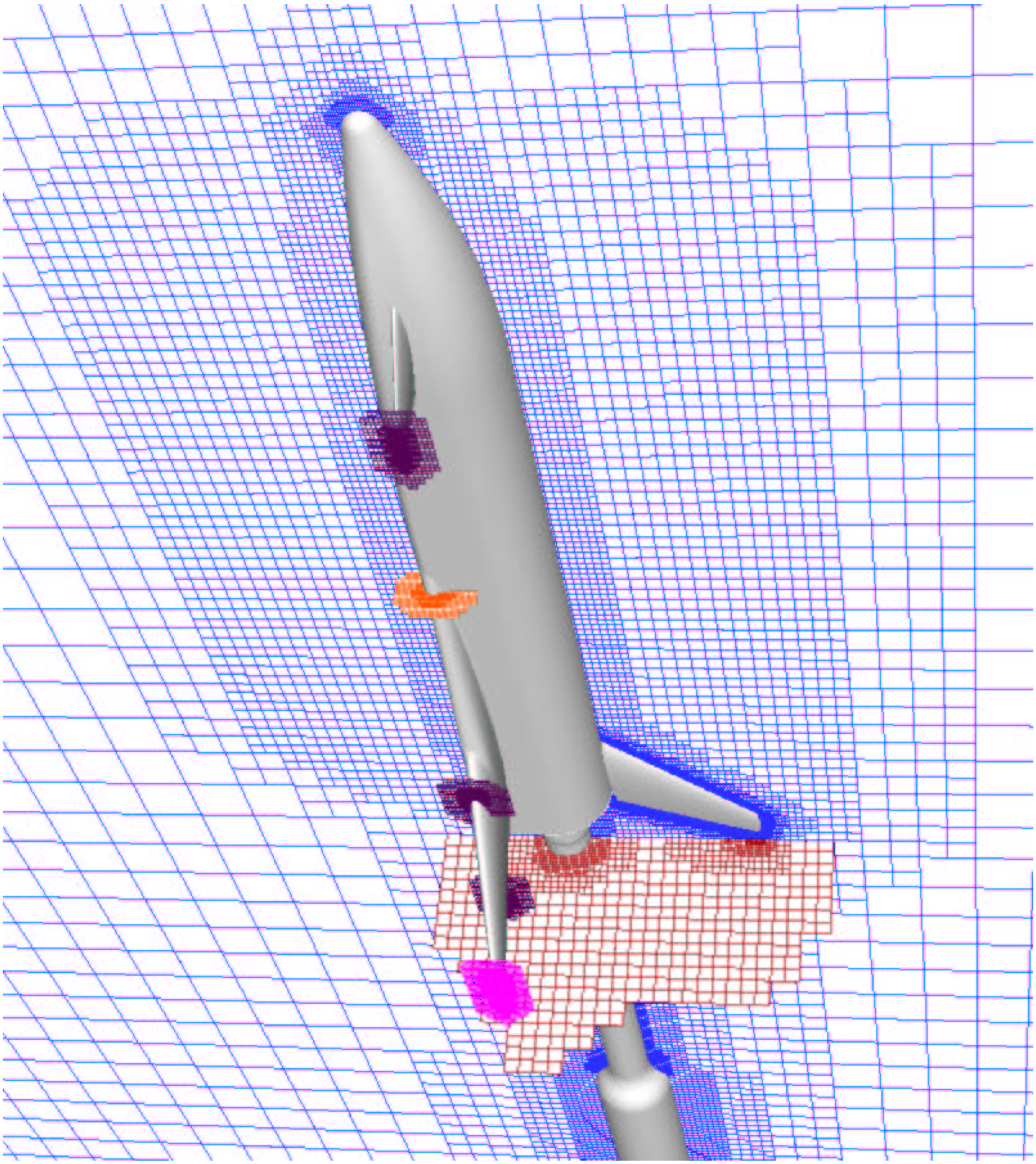
Department of Applied Mathematics and Statistics
University of California, Santa Cruz

Collaborators: Robert Gramacy (University of Cambridge)
Matt Taddy (UCSC)
William MacReady (formerly NASA Ames)

`http://www.ams.ucsc.edu/~herbie`

Outline

1. Motivating Example
2. Gaussian Process Models
3. Bayesian Treed GP Models
4. Limiting Linear Models
5. Adaptive Sampling
6. R code



Goal

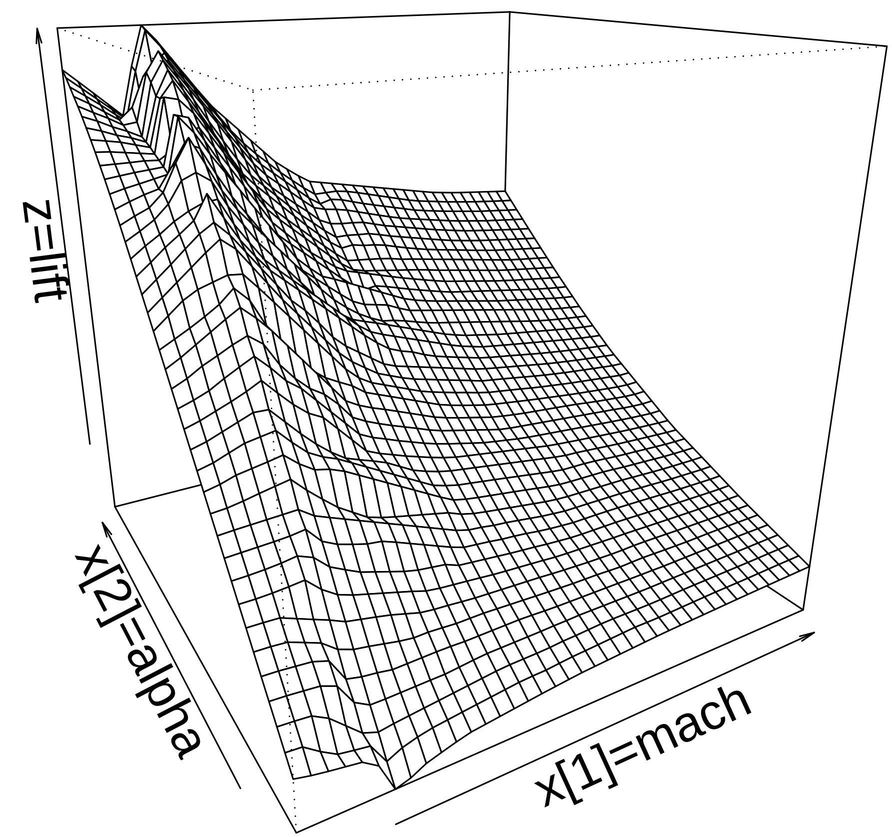
Develop a **response surface** for the model outputs of:

- lift
- drag
- pitch
- side-force
- yaw
- roll

for a variety of flight conditions (inputs):

- speed (Mach number)
- angle of attack (alpha)
- side slip angle (beta)

Estimated Surface



Overall Approach

Combine **modeling** and **sequential experimental design**.

- Start with an initial small run
- Fit a Bayesian model and estimate predictive uncertainty
- Choose new run locations based on uncertainty
- Attempt to put more effort in "interesting" regions

Gaussian Process Models

Traditional approach to modeling computer experiment output is a **Gaussian Process** (**GP**) (Sacks et al., 1989; Santner et al., 2003).

$$\mathbf{Z}(\mathbf{x}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}(\mathbf{x})$$

$\mathbf{Z}$ model outputs

$\mathbf{x}$ an arbitrary input value

$\mathbf{X}$ model inputs at all currently known data points

$\boldsymbol{\beta}$ linear trend coefficients

$\mathbf{W}$ mean zero spatial process

Spatial Correlation

W has covariance matrix **C** with elements

$$\mathbf{C}(\mathbf{x}_j, \mathbf{x}_k) = \sigma^2 \rho(\Delta), \text{ where } \Delta = \text{dist}(\mathbf{x}_j, \mathbf{x}_k)$$

$\rho(\Delta)$ is a correlation function, often the Matérn class or the power family applied to Euclidean distance

$$\rho(\Delta) = \exp\left\{-\frac{\Delta^{p_0}}{d}\right\}$$

can scale dimensions differently
can include a nugget effect

Drawbacks

- Scales poorly — matrix inversions are $O(n^3)$
- Strictly stationary — our data aren't
- Predictive error depends only on aggregate (not nearby) previously observed responses

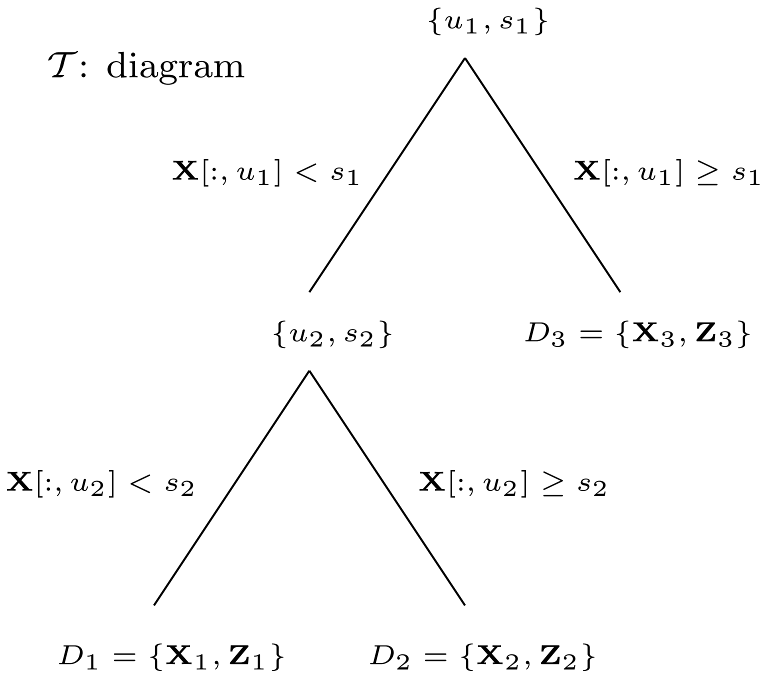
Solution: Partitioning

- Use a binary tree structure to **recursively partition** the space
  - Allow multiple splits per variable
- Tree process prior (Chipman et al., 1998)
- Fit a separate GP on each partition
- Fitting of tree structure and GPs is done **simultaneously** through MCMC
- Extension of partitioned linear regression model (Chipman et al., 2002)
- Nonstationarity achieved through partitioning

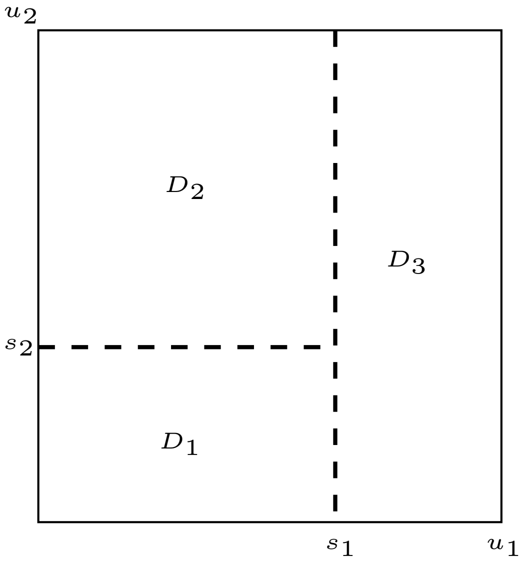
Tree Example

How a tree $\mathcal{T}$ recursively partitions the input space:

$\mathcal{T}$: diagram



$\mathcal{T}$: graphically



Fit with MCMC

Sample from the **joint posterior** of $(\mathcal{T}, \boldsymbol{\theta})$

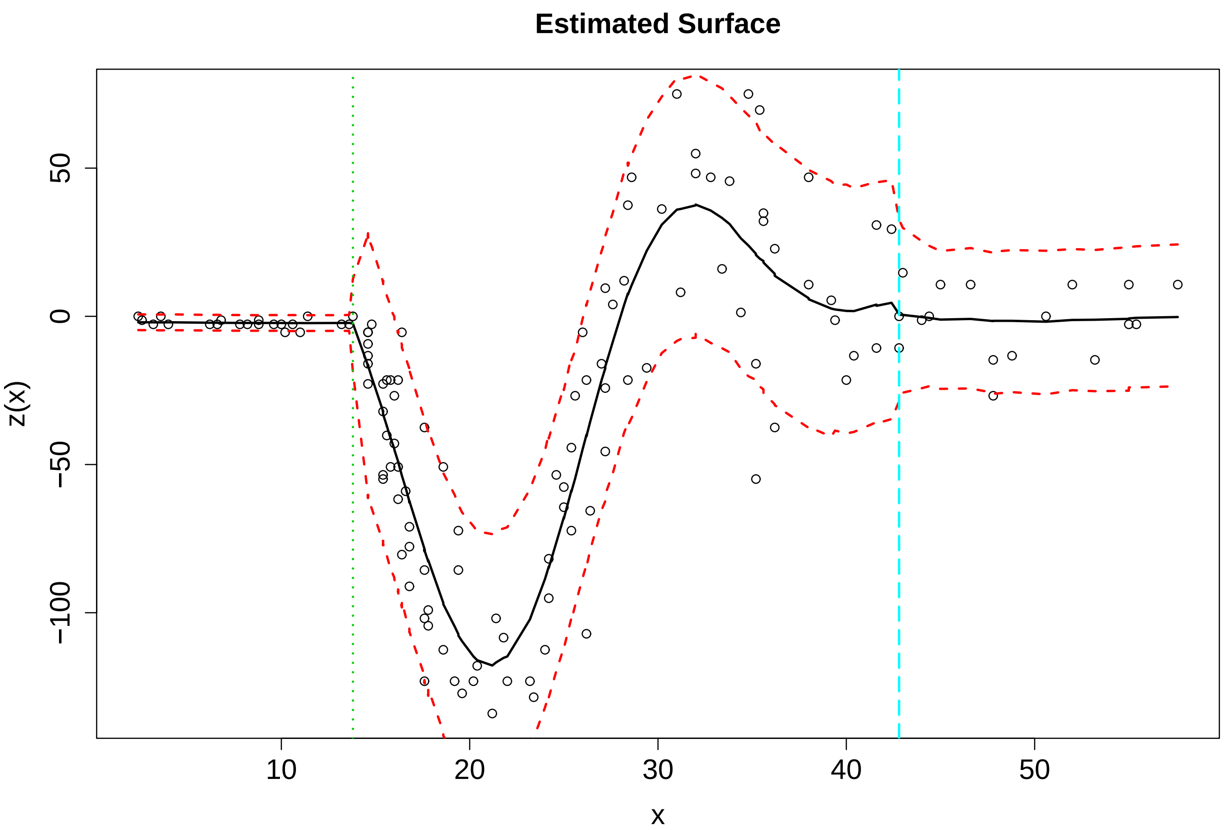
(Richardson & Green, 1997; Chipman et al., 2002)

- Average over $\mathcal{T}$ with **reversible-jump MCMC** (RJ-MCMC)
- Tree operations: **grow**, **prune**, **change**, **swap**, and **rotate**

Gibbs sampling for all within-partition parameters except the correlation, which requires Metropolis-Hastings

Motorcycle Accident Data

(Silverman, 1985)



Partly non-linear, non-stationary, heteroscedastic data

Limiting Linear Models

- Would like to be able to fit this dataset **accurately** and **efficiently**
  - Fit three distinct regions
  - Linear in first and third, non-linear in middle
  - Error variability differs by region
- A smooth fitted Gaussian Process can look essentially linear
- Combine with **treed partitioning** to fit either a full GP or a LLM in each of several partitions fit to the data
- Allow a region to jump between a smooth GP and a linear model
- Fit all parts of the model (parameters, GP/LLM, partitions) **simultaneously** with RJMCMC
- Results in a highly flexible yet computationally tractable non-stationary semi-parametric model.

Adaptive Sampling

Active Learning / Sequential Design of Experiments

- select future design sites to improve our knowledge (model)
- maximize some measure of **utility**
  - Kullback-Leibler distance between posterior predictive and prior predictive — equivalent to **minimizing predictive variance**
  - Choose next point as the one which has the largest predictive variance (MacKay, 1992) (ALM)

Issues

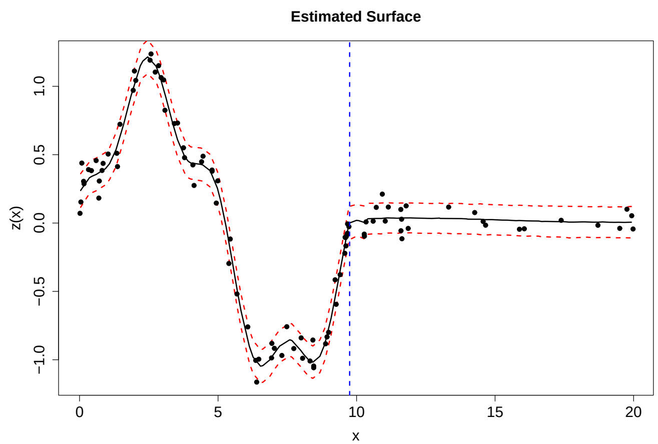
- Standard design approaches assume the model and its parameters are known
- Standard optimal designs tend to push points to the boundaries
- Searching over continuous space must be done approximately
- Need to select a list of multiple points
- Need to deal with pending data (experiments currently being run)

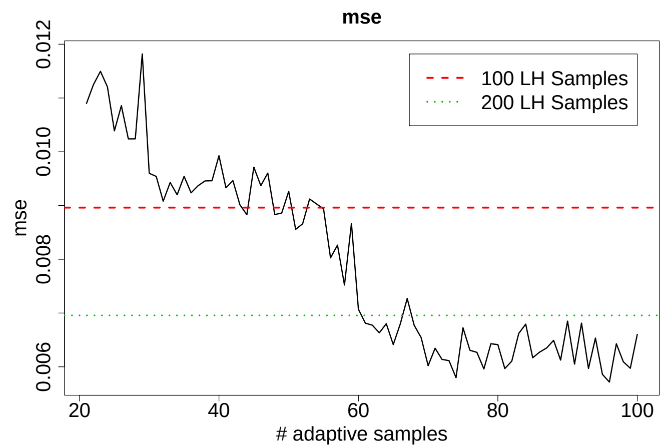
Iterative Algorithm

- Use MCMC to fit parameters and estimate variances using currently known data plus fitted values at pending locations
- Create prioritized list of new design points based on predictive variance
- Incorporate results of new experimental runs
- Clear or prune tree structure and repeat

Adaptive Sampling Demo

$$z(x) = \begin{cases} \sin\left(\frac{\pi x}{5}\right) + \frac{1}{5}\cos\left(\frac{4\pi x}{5}\right) & x < 10 \\ 0 & \text{otherwise} \end{cases}$$



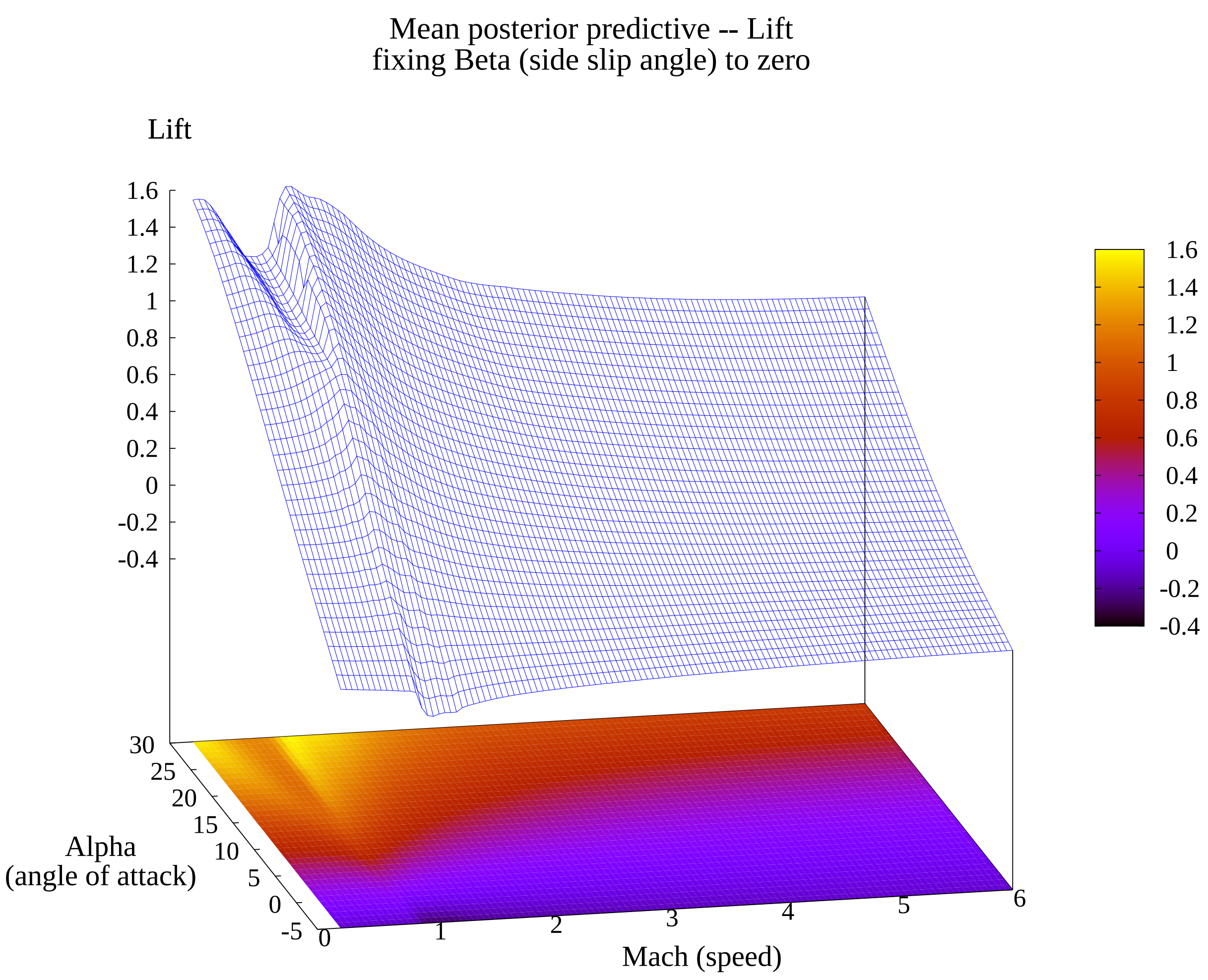


- MSE comparison to LH designs with **100** and **200** samples

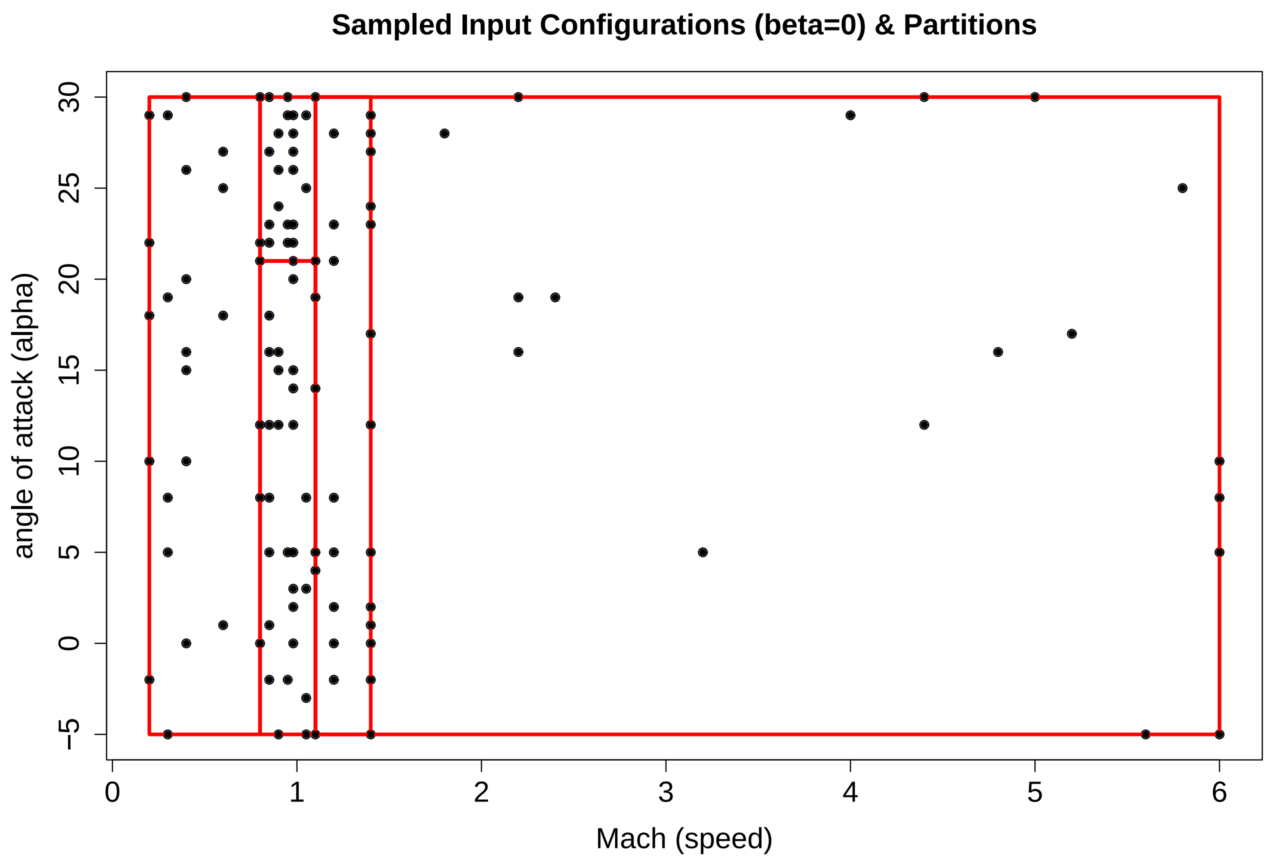
Rocket Booster Example

- 3 inputs, 6 outputs
- Each sample required 5-20 hours computing time
- Non-stationary
- Fit independent treed GP for each response, use standardized average of predictive variability

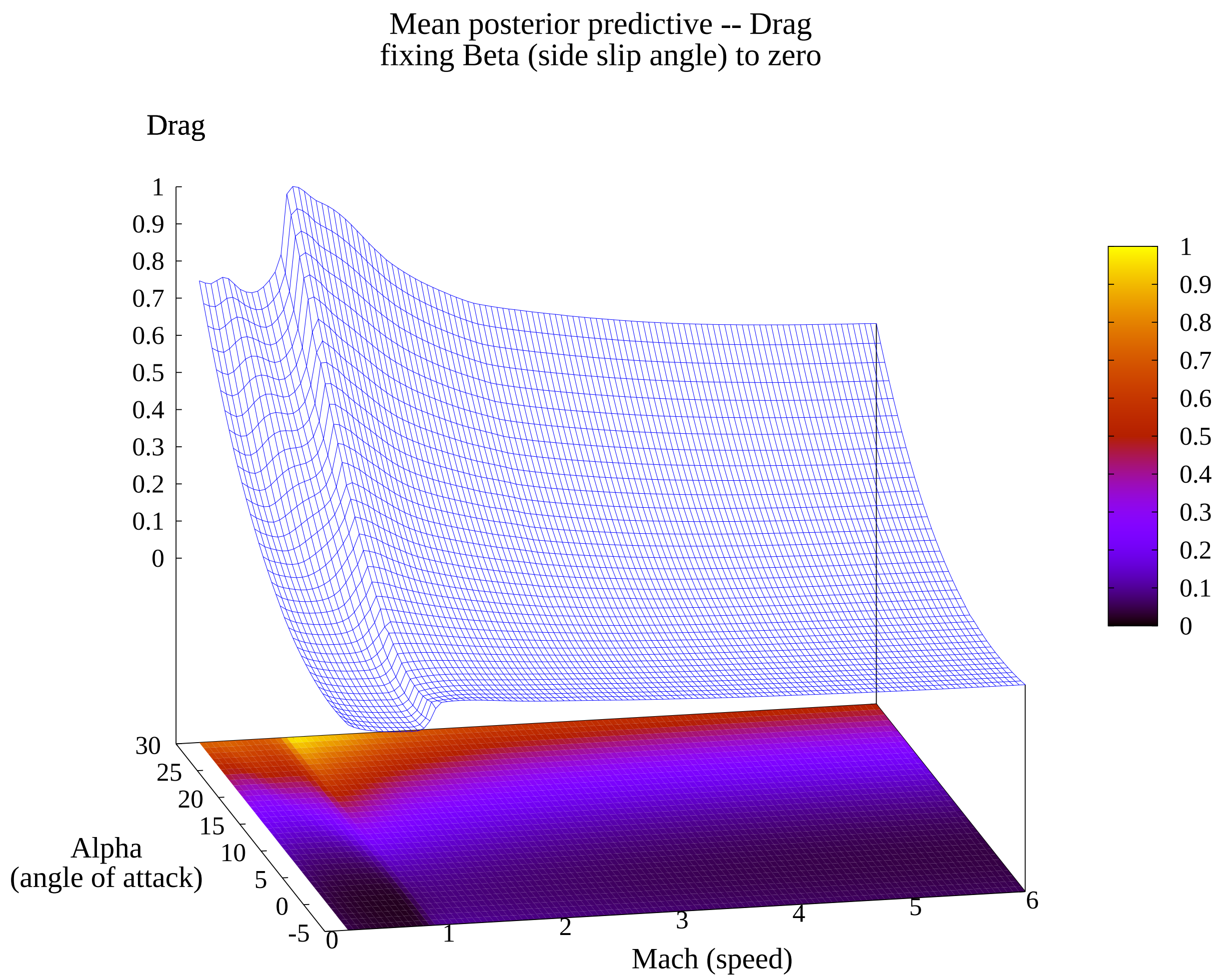
Adaptive Sampling on LGBB: Lift



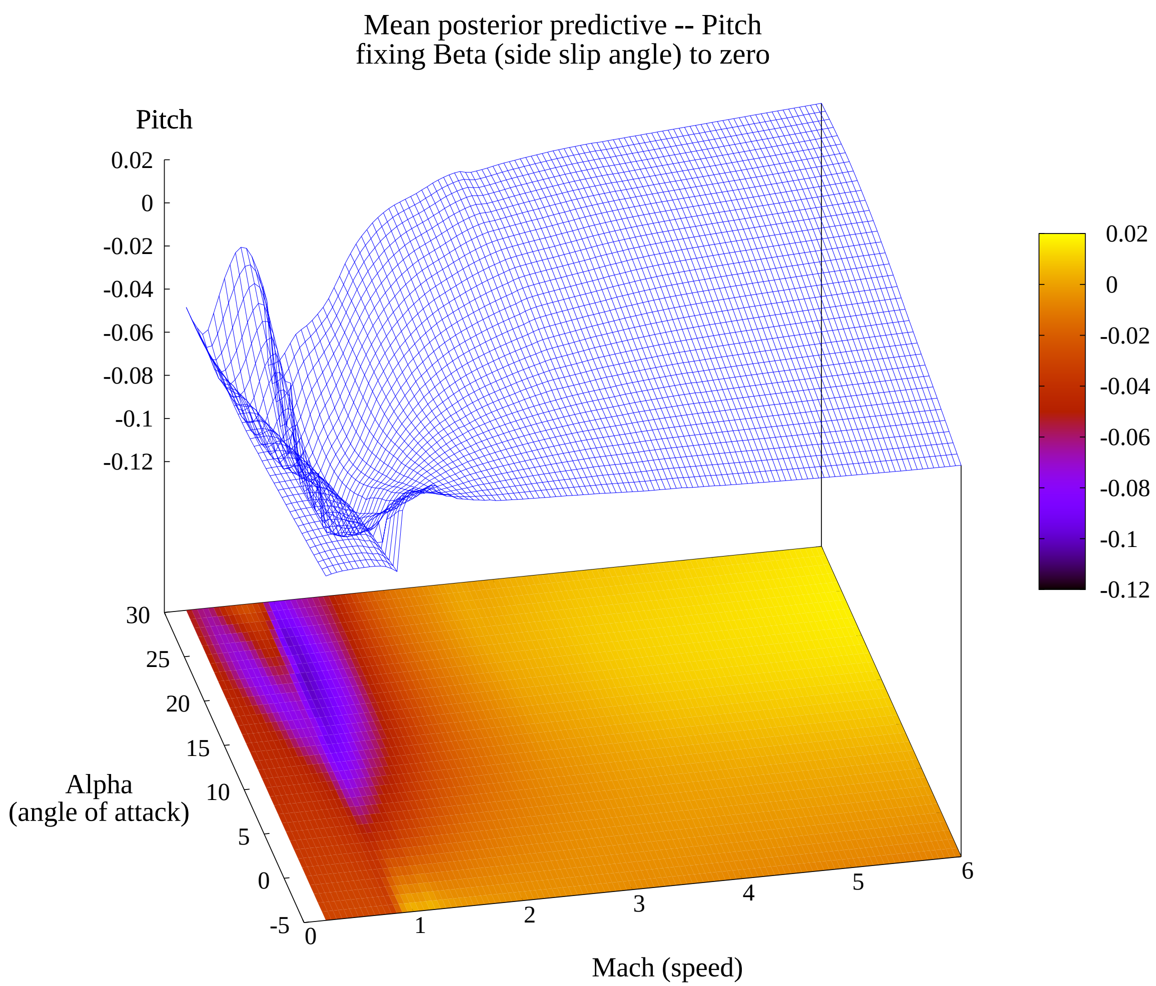
Adaptive Sampling on LGBB: Lift



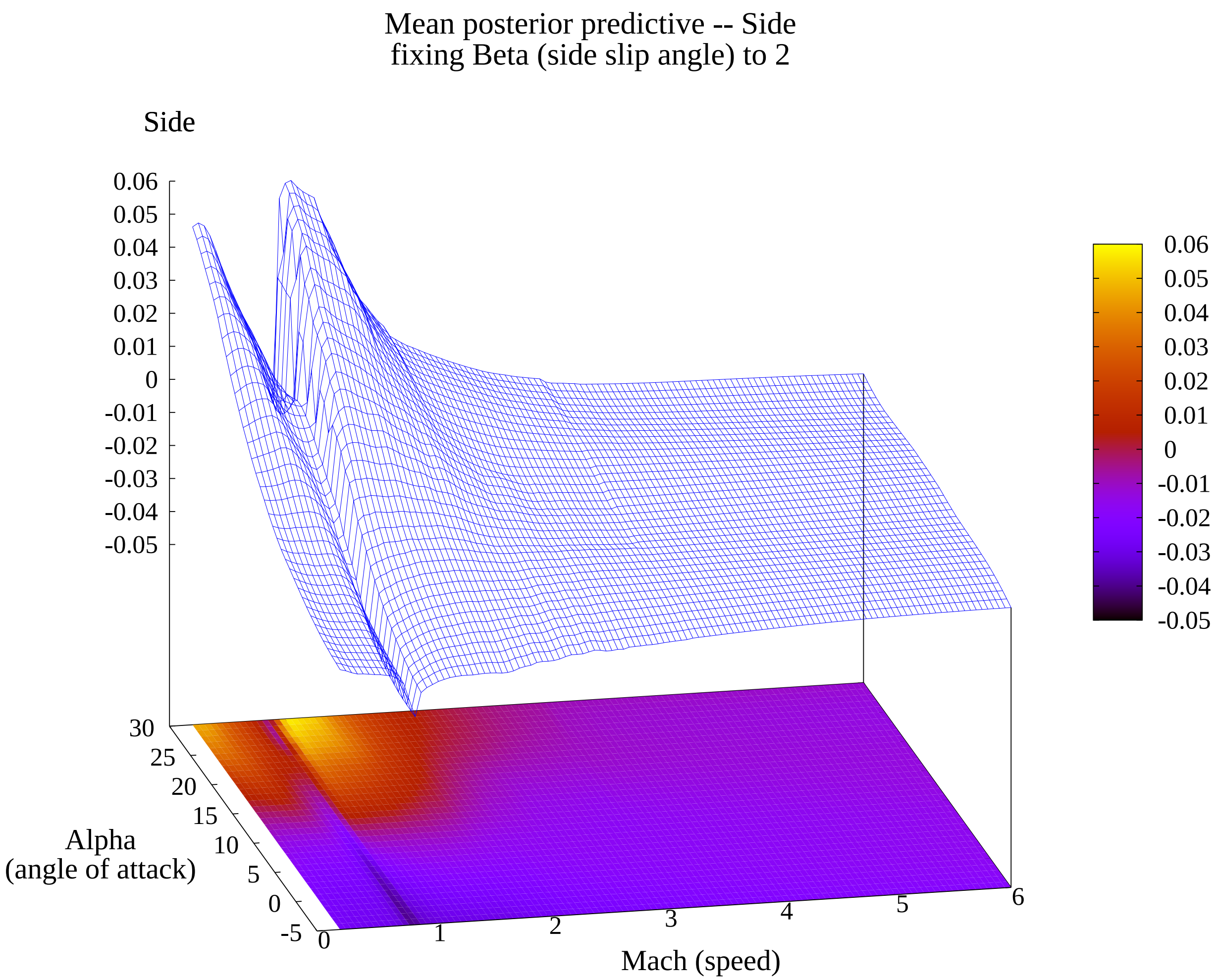
Adaptive Sampling on LGBB: Drag



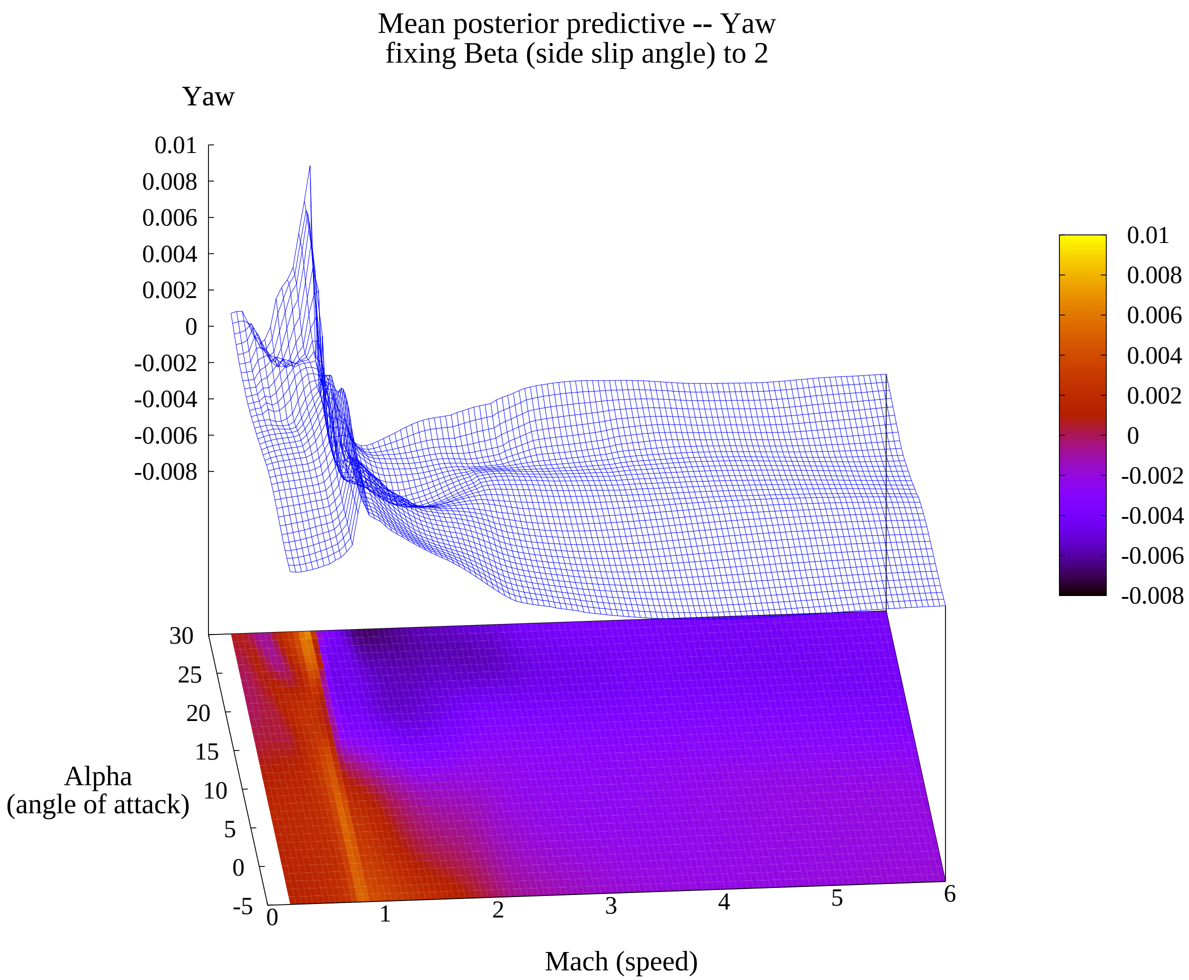
Adaptive Sampling on LGBB: Pitch



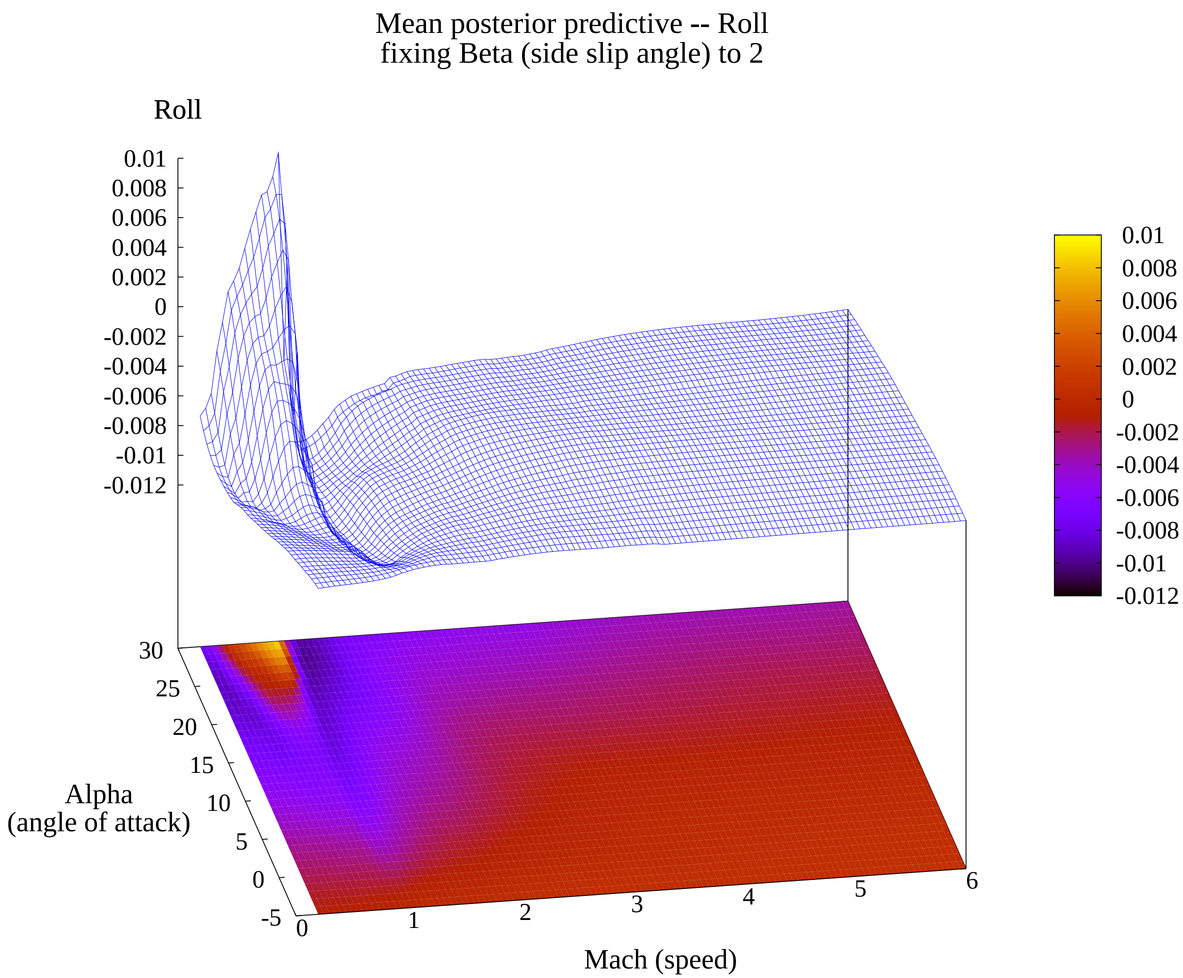
Adaptive Sampling on LGBB: Side Force



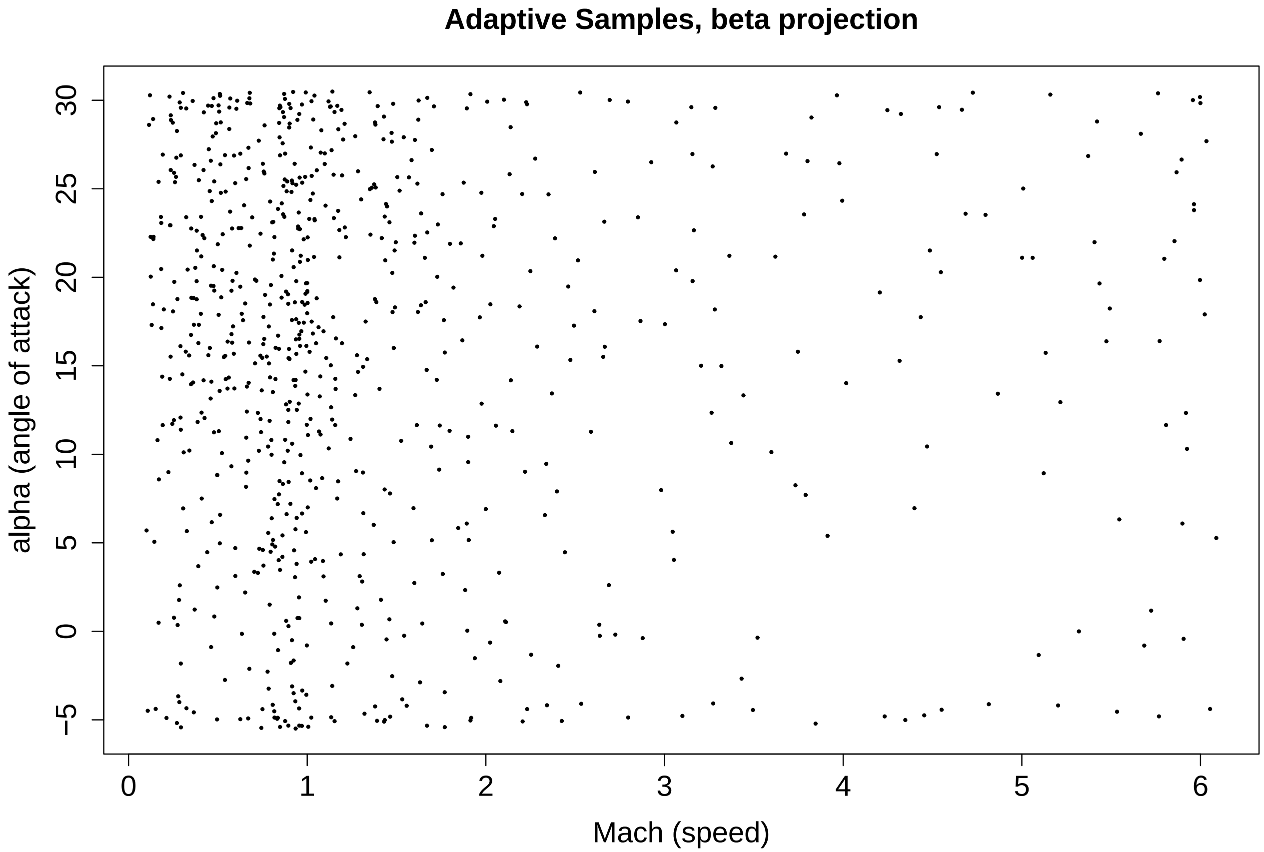
Adaptive Sampling on LGBB: Yaw



Adaptive Sampling on LGBB: Roll



Adaptive Sampling on LGBB



750 adaptive samples, compared to more than **3250** at NASA

tgp R library

Available at

http://www.cran.r-project.org/src/contrib/Descriptions/tgp.html

Or just get within R with

```
install.packages("tgp")
```

References

Bernstein, D. (2005). *Matrix Mathematics*. Princeton, NJ: Princeton University Press.

Chipman, H., George, E., and McCulloch, R. (1998). "Bayesian CART Model Search (with discussion)." *Journal of the American Statistical Association*, 93, 935–960.

— (2002). "Bayesian Treed Models." *Machine Learning*, 48, 303–324.

Cressie, N. A. C. (1993). *Statistics for Spatial Data, revised edition*. John Wiley and Sons.

Gramacy, R. B., Lee, H. K. H., and Macready, W. (2004). "Parameter Space Exploration With Gaussian Process Trees." In *ICML*, 353–360. Omnipress & ACM Digital Library.

MacKay, D. J. C. (1992). "Information-based Objective Functions for Active Data Selection." *Neural Computation*, 4, 4, 589–603.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). "Design and Analysis of Computer Experiments." *Statistical Science*, 4, 409–435.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. New York, NY: Springer-Verlag.

Silverman, B. W. (1985). "Some Aspects of the Spline Smoothing Approach to Non-Parametric Curve Fitting." *Journal of the Royal Statistical Society Series B*, 47, 1–52.