Module 11: Gaussian Process Regression

DAV-6300-1: Experimental Optimization

David Sweet // 20241121

Review: A/B Test

- Goal: Accept or reject B
- Design: $N \geq \left(\frac{2.5\hat{\sigma}_\delta}{PS}\right)^2$
- Measure: Replicate (reduce variance), Randomize (reduce bias)
- Analyze:

Criterion 1: $\delta > 1.6se \;\; (t > 1.6)$

Criterion 2: $\delta > PS$

Review: Thompson sampling

- Allocate observations to arms in proportion to the probability each arm is best
  - $p_{\text{arm}} \propto p_{\text{best}}$
- Stop when $\max\{p_{\text{best}}\} > 0.95$

Review: Response Surface Methodology

- Surrogate: Model (regression)
  - Maps parameters, x, to measurements, y
- Analogy
  - $E[BM]$ is to observation y
  - as response function, $f(x)$, is to surrogate, $y(x)$

Key Terms

- Surrogate (again)
- Gaussian Process
- Gaussian Process Regression (GPR)
- Non-parametric
- Aleatoric (measurement) & epistemic (model) uncertainty

Gaussian Process Regression

A modern, powerful surrogate

- Recall, RSM uses linear model
  - $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
  - Engineer decides regressors
  - Engineer fits & inspects model

Gaussian Process Regression

A modern, powerful surrogate

- Recall, RSM uses linear model
  - $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
  - Engineer decides regressors
  - Engineer fits & inspects model
- GPR uses non-parametric model
  - "Fancy" KNN
  - No regressors, just data
  - GPR just works
  - GPR used in Bayesian Optimization (BO)

Gaussian Process Regression

A modern, powerful surrogate

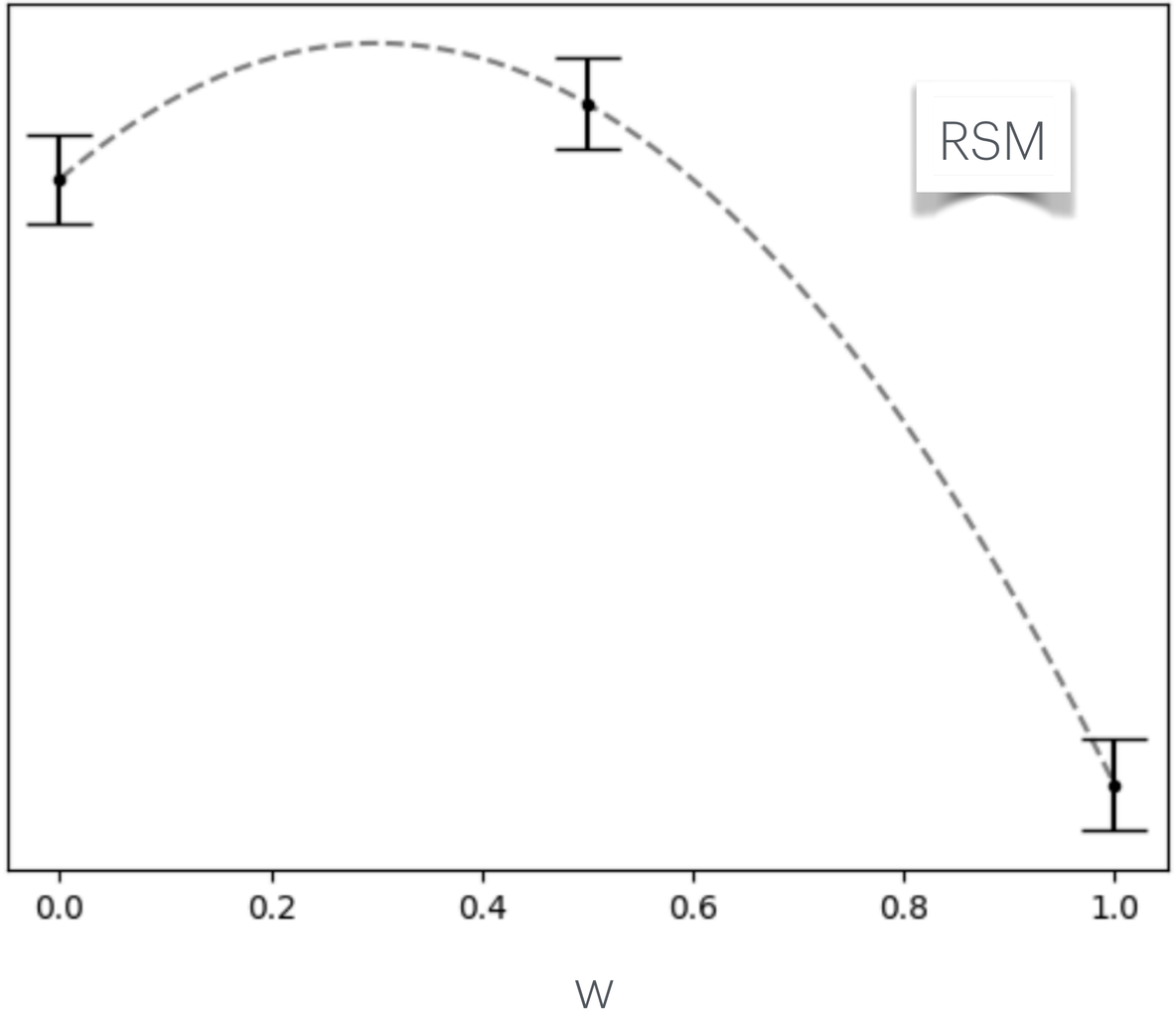
- RSM
  - Rigid model form: one hump
  - Best for few dimensions
  - Models only $\mu(x)$
  - "Statistics"

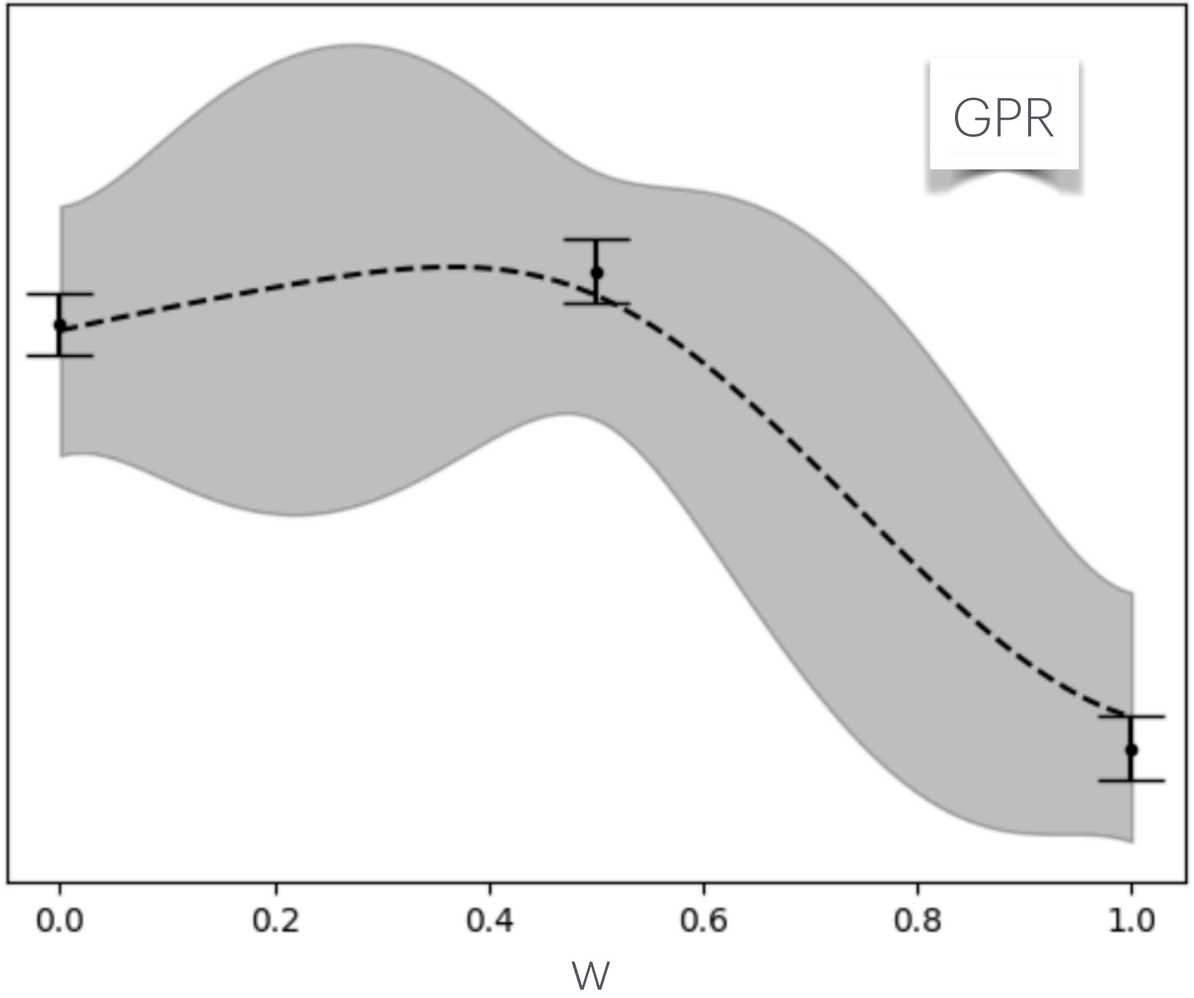
- GPR uses non-parametric model
  - Flexible; any shape
  - Fine for any number of dimension
  - Models $\mu(x)$ and $se(x)$
  - "Machine Learning"

Hard

Easy

Gaussian Process Regression





GP Model

- RSM
  - $y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
  - $\varepsilon \sim \mathcal{N}(0, se^2)$
- Same as $y(x) \sim \mathcal{N}(\mu(x), se^2)$
  - $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

- GPR extends
  - $y(x) \sim \mathcal{N}(\mu(x), se^2\mathbf{(x)})$
- se now depends on x

$\mu(x)$

- Want $\mu(x)$ as weighted avg. of all y_i's
- How similar are nearby measurements?
- Nearness: $d(x, x') = \|x - x'\| = \sqrt{\sum_i (x_i - x_i')^2}$

Euclidean distance

- Similarity: $e^{-d(x,x')^2/(2s^2)}$

RBF/Squared exponential Kernel

$\mu(x)$

- kernel function: $k(x, x') = e^{-d(x,x')^2/(2s^2)}$
- Kernel matrix, all pairs of parameters: $(K_{xx})_{ij} = k(x_i, x_j)$
- Kernel vector, estimate at x: $(K_x)_i = k(x, x_i)$

$$\mu(x) = K_x^T(K_{xx} + se_0^2 I)^{-1}\mathbf{y}$$

- $\mathbf{y}$ is vector, y_i, of all measurements
- se_0 is standard error of all y_i

See Appendix C of
Experimentation for Engineers

$\mu(x)$

$$\mu(x) = \underbrace{K_x^T(K_{xx} + se_0^2 I)^{-1}}_{\text{weights}}\mathbf{y}$$

- $\mu(x)$ is weighted avg. of y's
- Weights depend on kernel values, on distances between x's

$se(x)$

- se is similar

$$se^2(x) = 1 - K_x^T(K_{xx} + se_0^2 I)^{-1} K_x$$

- se_0^2 is measurement noise
- se_0^2 constant, common to all y's
- $se(x)$ depends only on x_i

Independent of measured BM

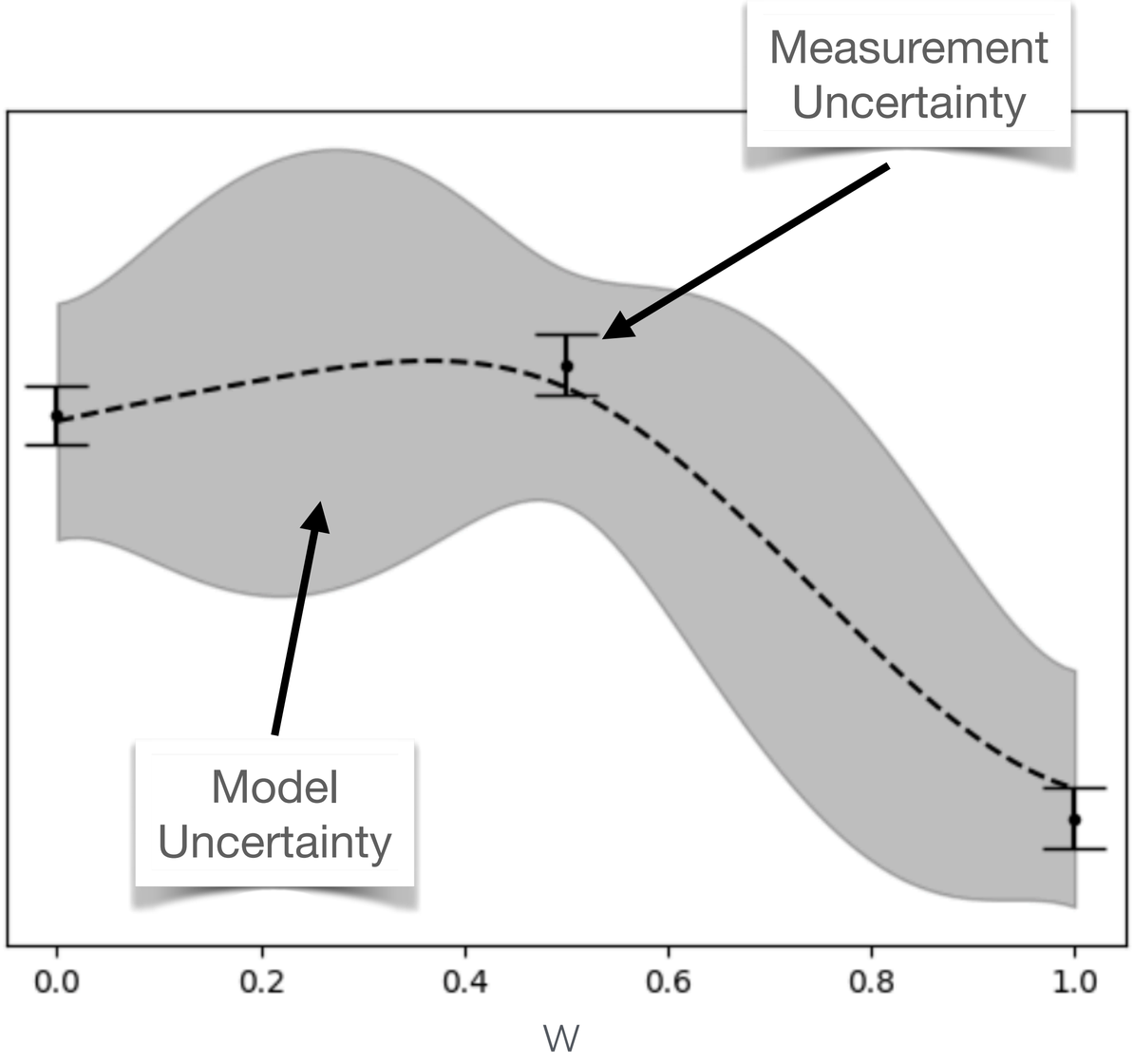
$se(x)$

$$se^2(x) = 1 - K_x^T(K_{xx} + se_0^2 I)^{-1} K_x$$

- se_0^2 is *aleatoric uncertainty* — measurement uncertainty
  - The familiar one
- $K_x^T K_{xx}^{-1} K_x$ is *epistemic uncertainty* — model uncertainty
  - Farther from measurements, greater uncertainty

$se(x)$

- Measurement uncertainty
  - Error bars
  - Decrease by increasing N
- Model uncertainty
  - Gray areas
  - Decrease by measuring a new parameter value



Computation

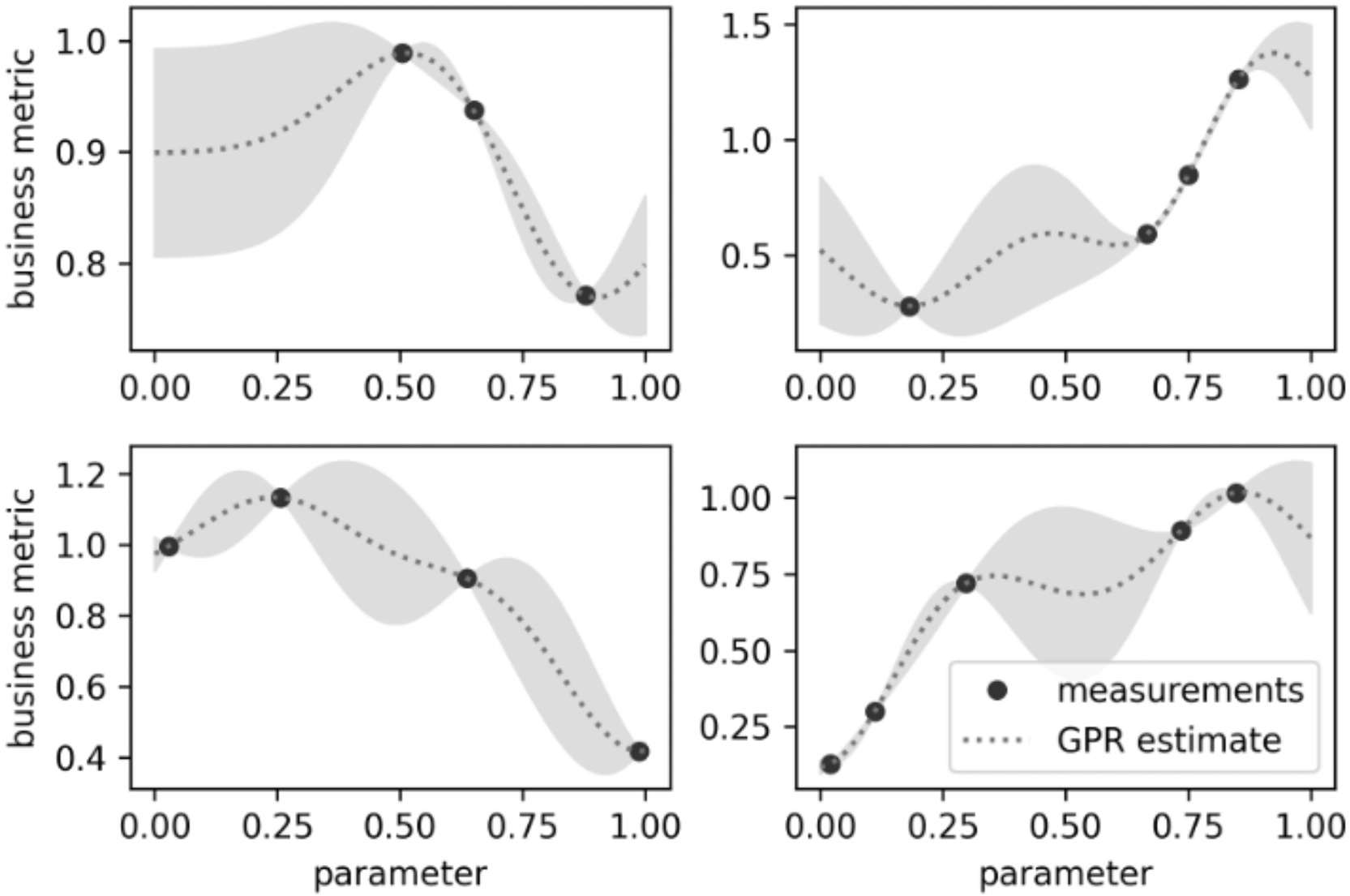
- Uses all $O(N^2)$ distances — memory hog
- Inverts K matrix, $O(N^3)$ — slow
- N is number of measurements
  - GPR good when N small
  - Experiments try hard to keep N small

larger N ==> more expensive

Kernels

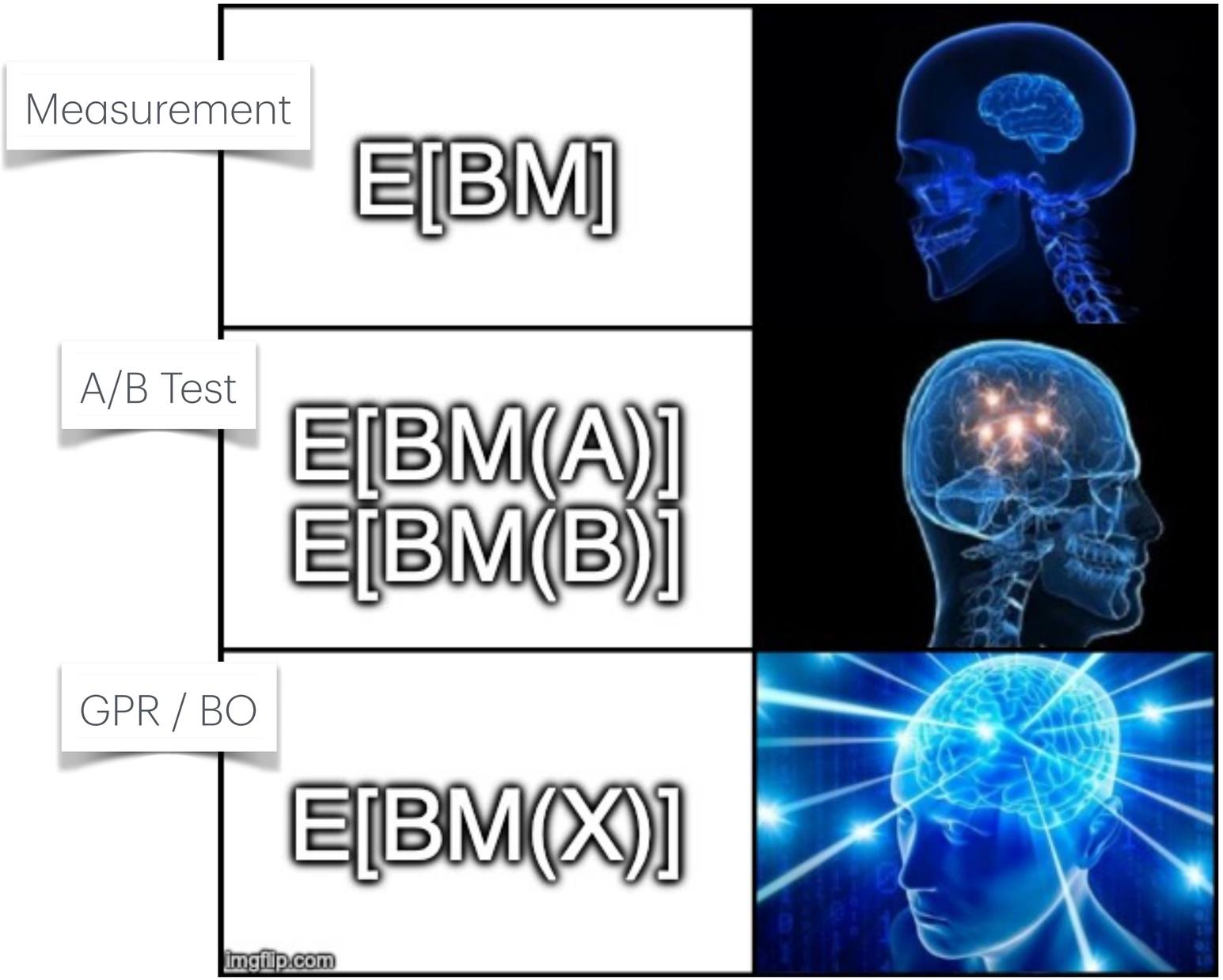
- Kernel function is part of model architecture
- Many kernel functions available
  - Localized, like RBF / Gaussian
  - Periodic
  - Nearness of long strings (molecular discovery)
  - Nearness of images

Examples


business metric
1.0
0.9
0.8
0.00
0.25
0.50
0.75
1.00
1.5
1.0
0.5
business metric
1.2
1.0
0.8
0.6
0.4
parameter
1.00
0.75
0.50
0.25
measurements
GPR estimate
parameter

Progression

- If x is indicator:
  - $x = 0$ if A
  - $x = 1$ if B
- ...then GPR models an A/B test



What puts the G in GPR?

And how is it a "process"?

- Model each value $y(x)$ as a Gaussian distribution
- Model any collection of $\{y(x)\}$ as a multivariate Gaussian distribution
  - x is continuous, so really an infinite-dimension Gaussian distribution
- First considered as $y(t)$, where t is time. A process is something that changes over time. A Gaussian process is one where y has a Gaussian distribution that changes over time. Ex: a Brownian motion (continuous random walk)
- Change t to x and you have a machine learning tool, Gaussian process regression

Summary

- GPR models $\mu(x)$ and $se(x)$
- $se(x)$ models both aleatoric (measurement) and epistemic (model) uncertainties
- Non-parametric; no betas, like KNN
  - Reads all measurements for every new estimate
- Slow, but very good for experimentation, where N is small
- GPR used as surrogate in Bayesian optimization