Bayesian Optimization of Integrated Response Surfaces

Saul Toscano-Palmerin, Peter Frazier
Operations Research & Information Engineering, Cornell

Optimization of **expensive integrands:**

$$\max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \sum_{w=1}^{n} F\left(x, w\right) p(w)$$

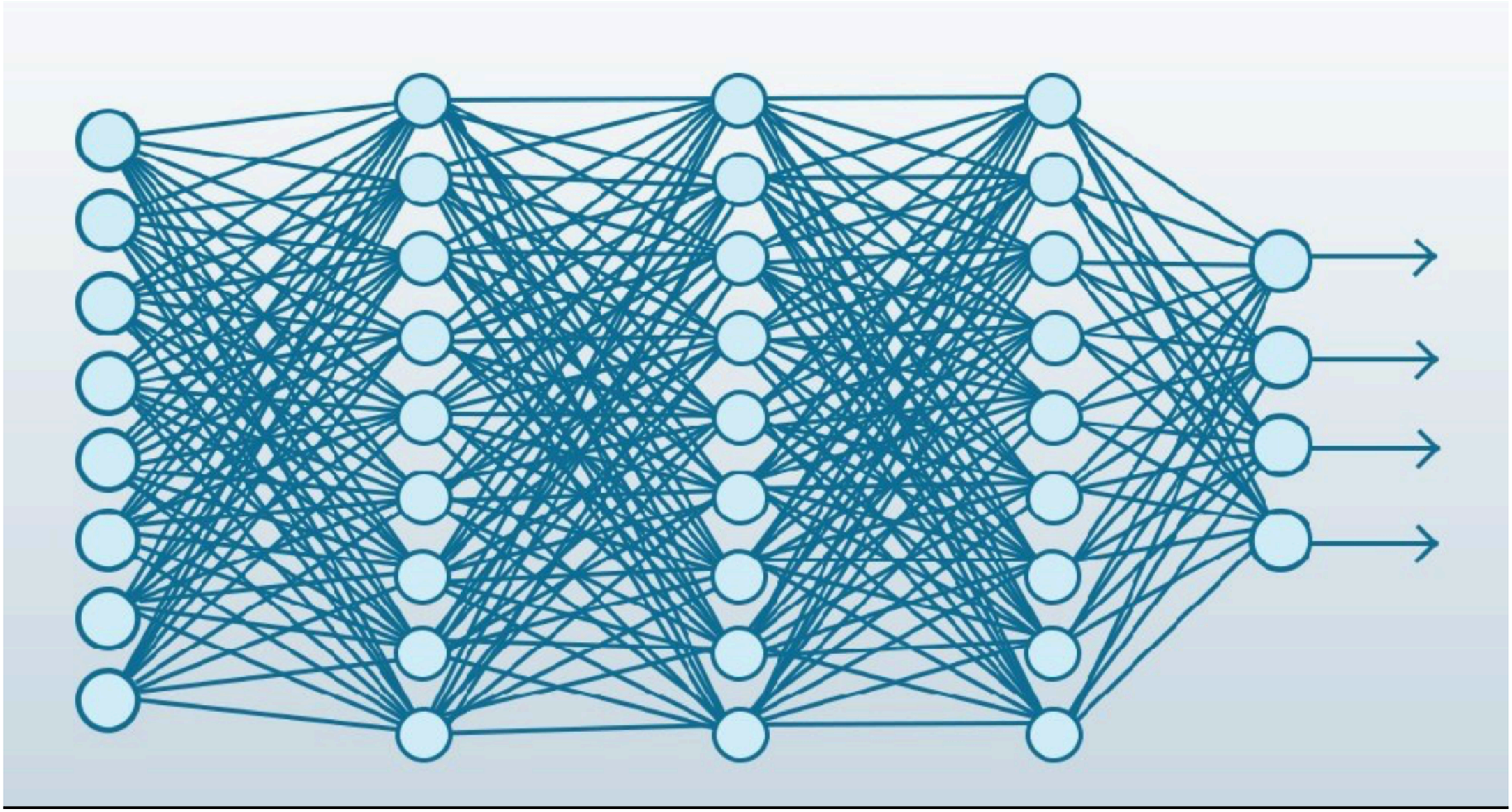
$$\max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \int F\left(x, w\right) p(w) dw$$

MULTI-TASK LEARNING

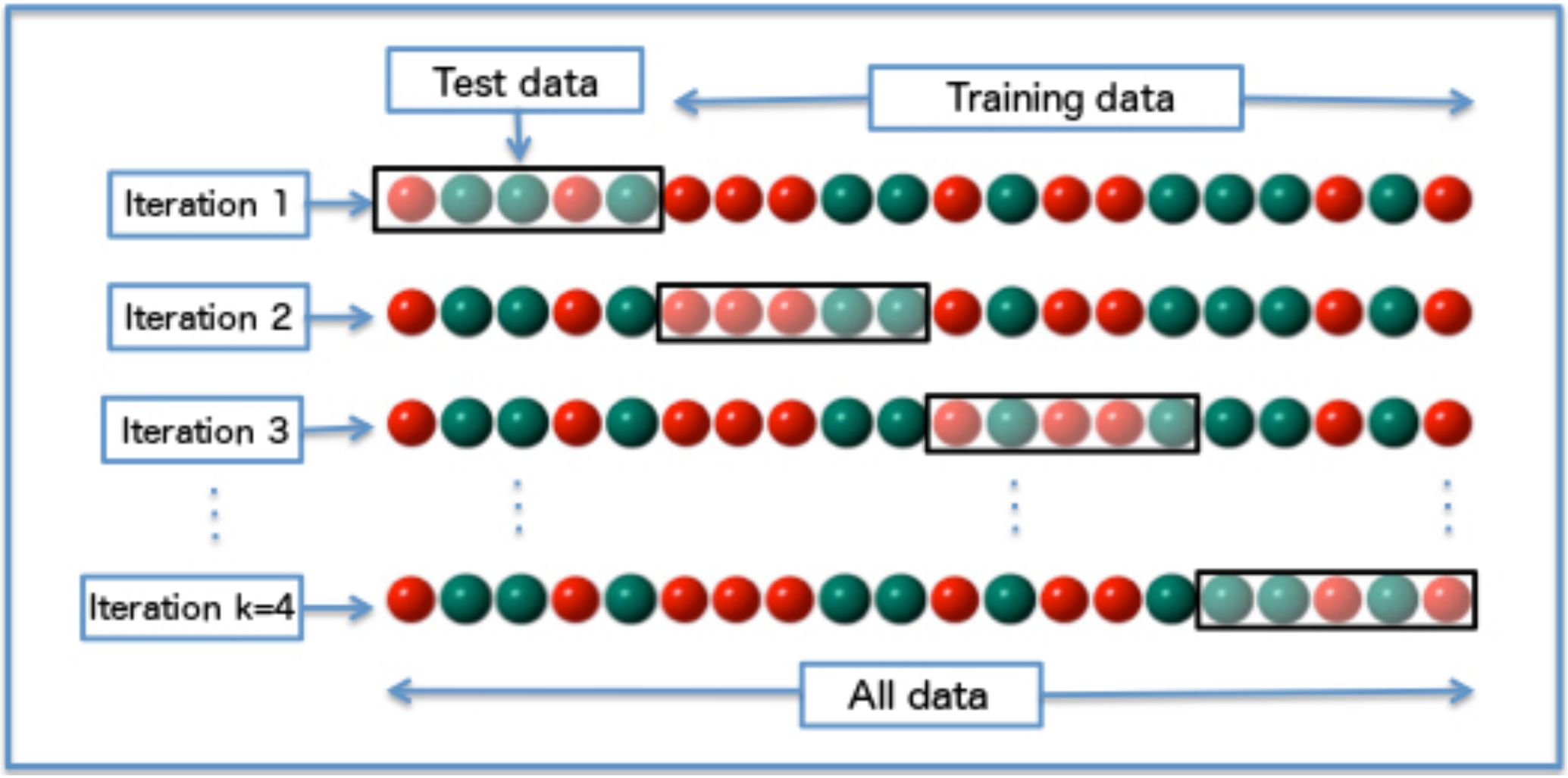
SIMULATION OPTIMIZATION

DESIGN OF EXPERIMENTS

Neural Network Design Using Cross-Validation



Neural Network Design Using Cross-Validation

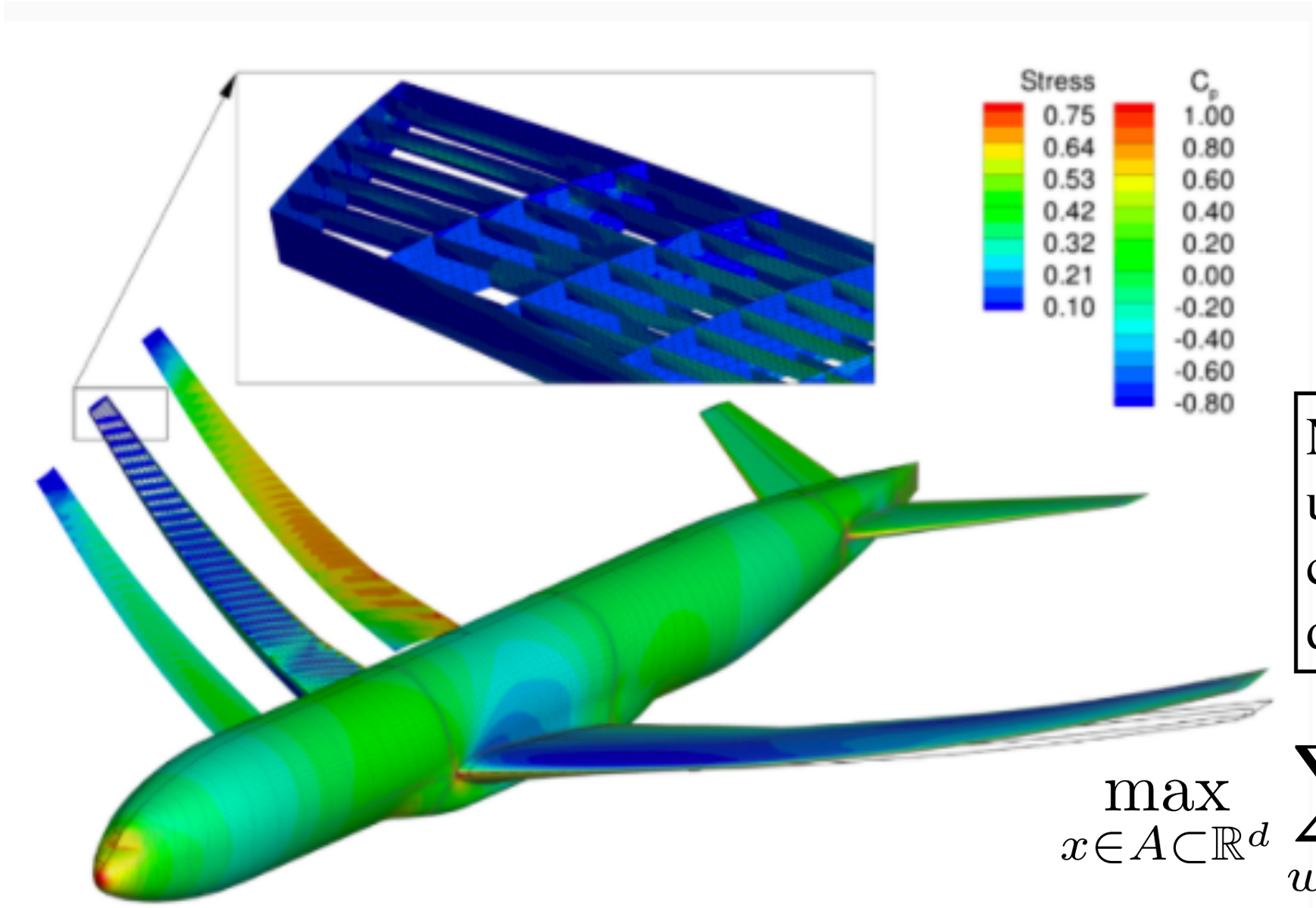


$$\max_{x \in A \subset \mathbb{R}^d} \sum_{w=1}^{n} F(x, w)\, p(w)$$

1/4

Negative error of the NN with hyperparameters x evaluated on the w-th dataset

Optimal Aircraft Design Under Different Flight Conditions



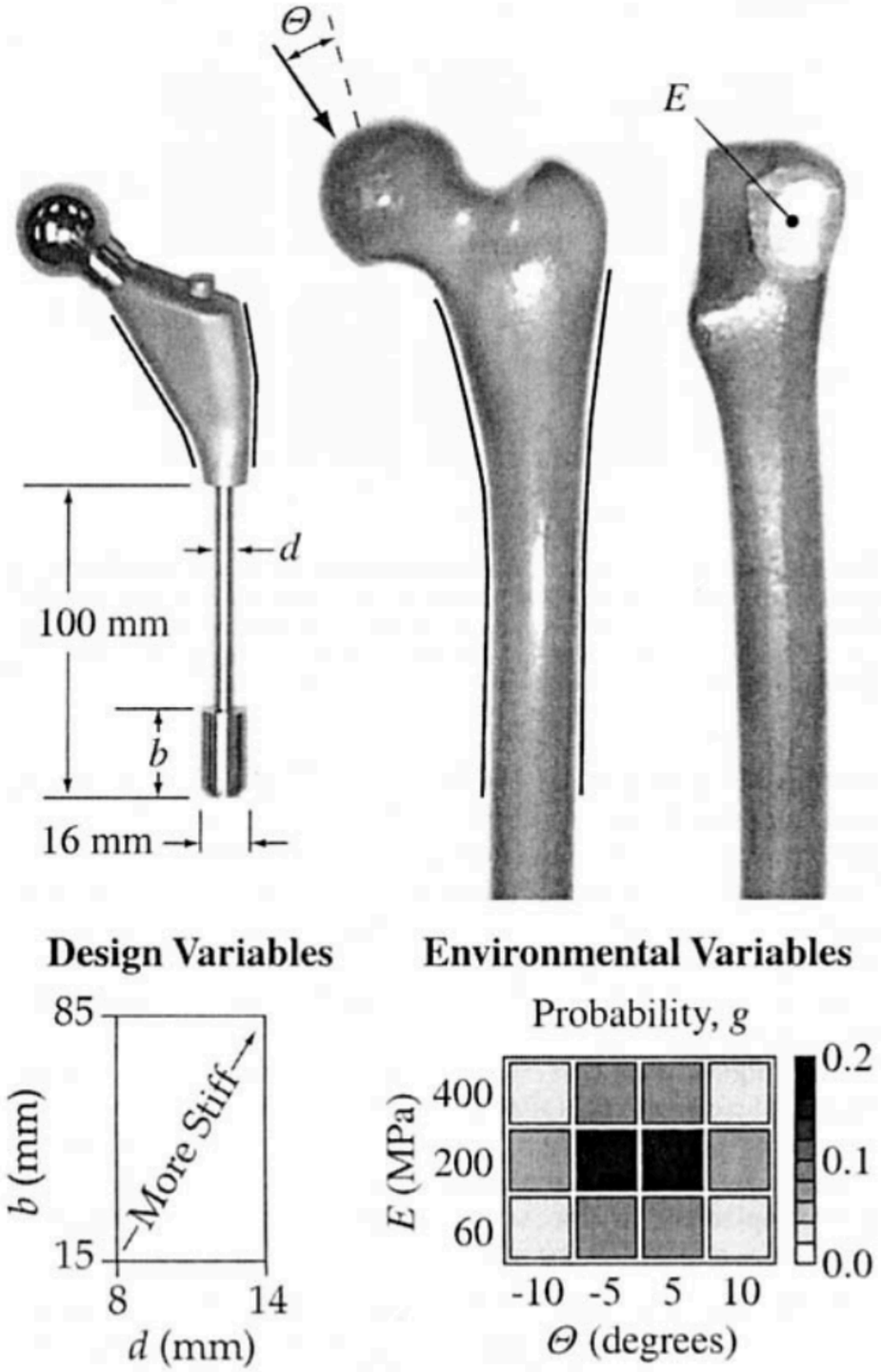
Liem et al., 2015

Negative fuel burn under flight conditions w and design of aircraft x

$$\max_{x \in A \subset \mathbb{R}^d} \sum_{w=1}^{n} F(x, w)\, p(w)$$

Probability of flight conditions w

Design of Experiments Under Random Environmental Conditions



Chang et al., 1999

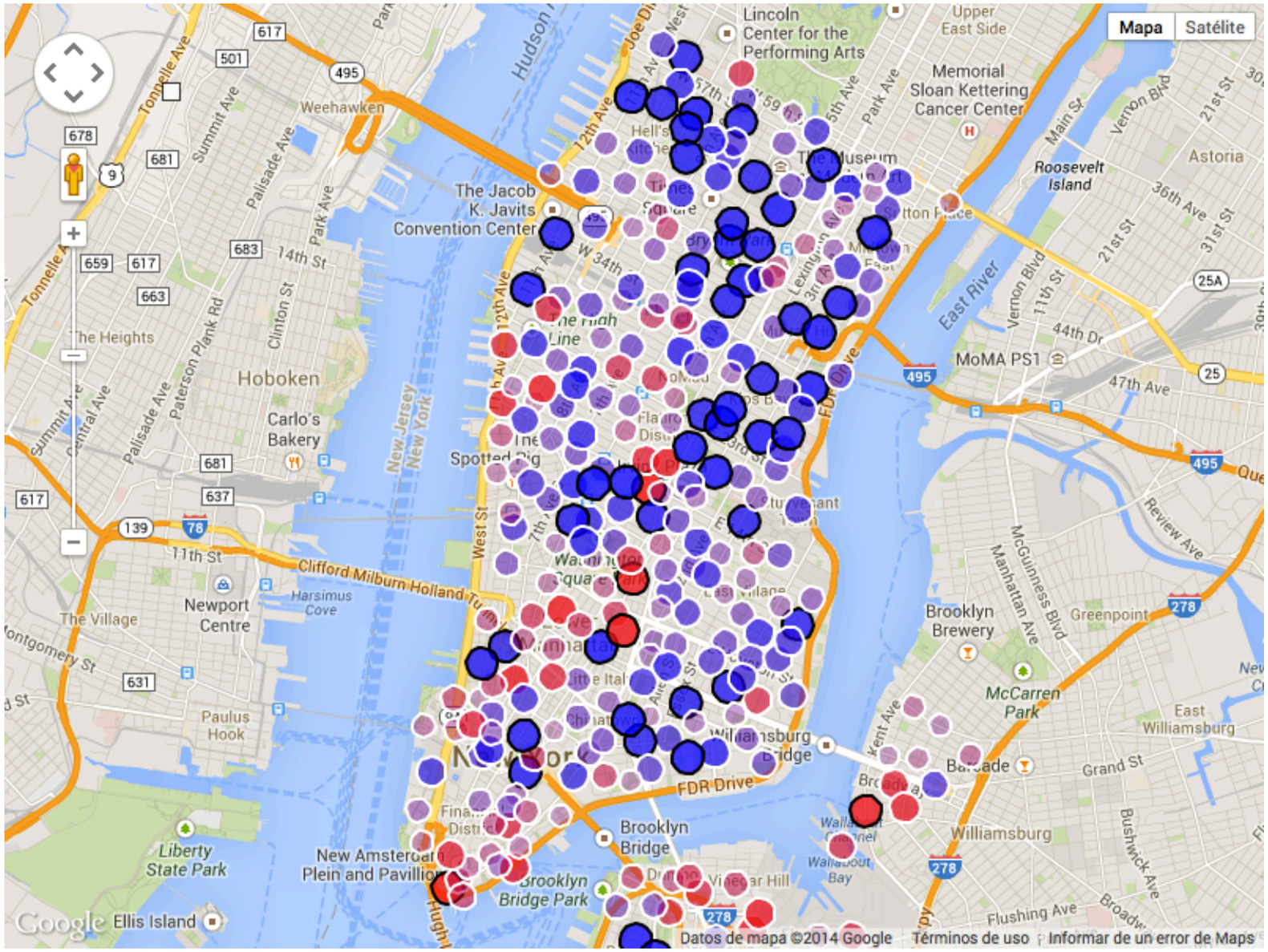
Optimal bone stress of a hip prosthesis

Probability of patient conditions w

$$\max_{x \in A \subset \mathbb{R}^d} \sum_{w=1}^{n} F(x, w)\, p(w)$$

Negative bone stress of design x under patient conditions w

Noise Reduction of Agent-Based Models



Poisson mixture density

$$\max_{x \in A \subset \mathbb{R}^d} \int F(x, w)\, p(w) dw$$

Negative expectation of dissatisfied trips given that the overall trip demand is w, and the allocation of bikes is x

Noise Reduction in Inventory Problems

Density of a multivariate Gumbel distribution

$$\max_{x \in A \subset \mathbb{R}^d} \int F(x, w)\, p(w) dw$$

Profit given the inventory x, and the sum of a multivariate Gumbel distribution w

The General Problem

- We want to optimize:

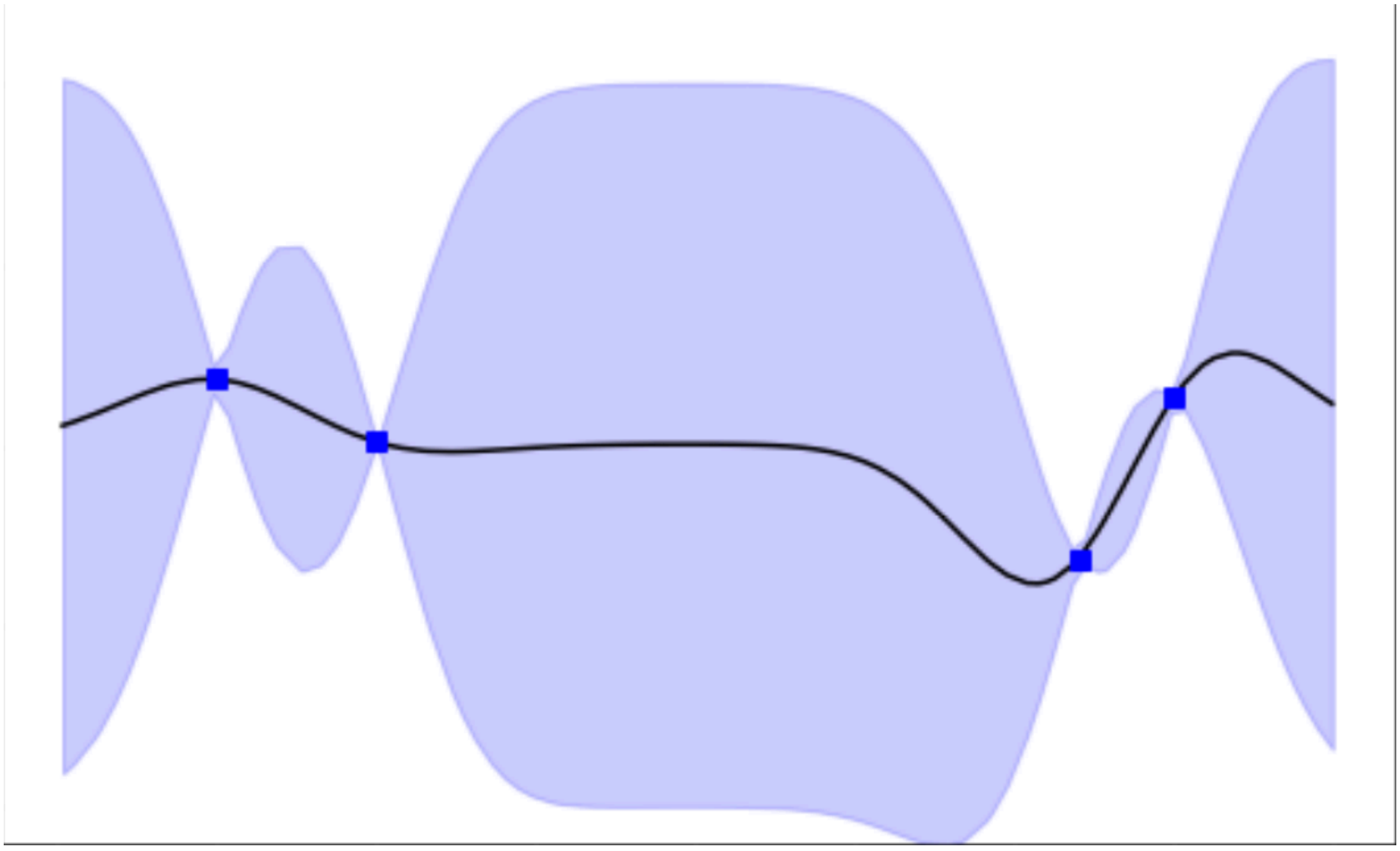
$$\max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \sum_{w=1}^{n} F\left(x, w\right) p(w)$$

or

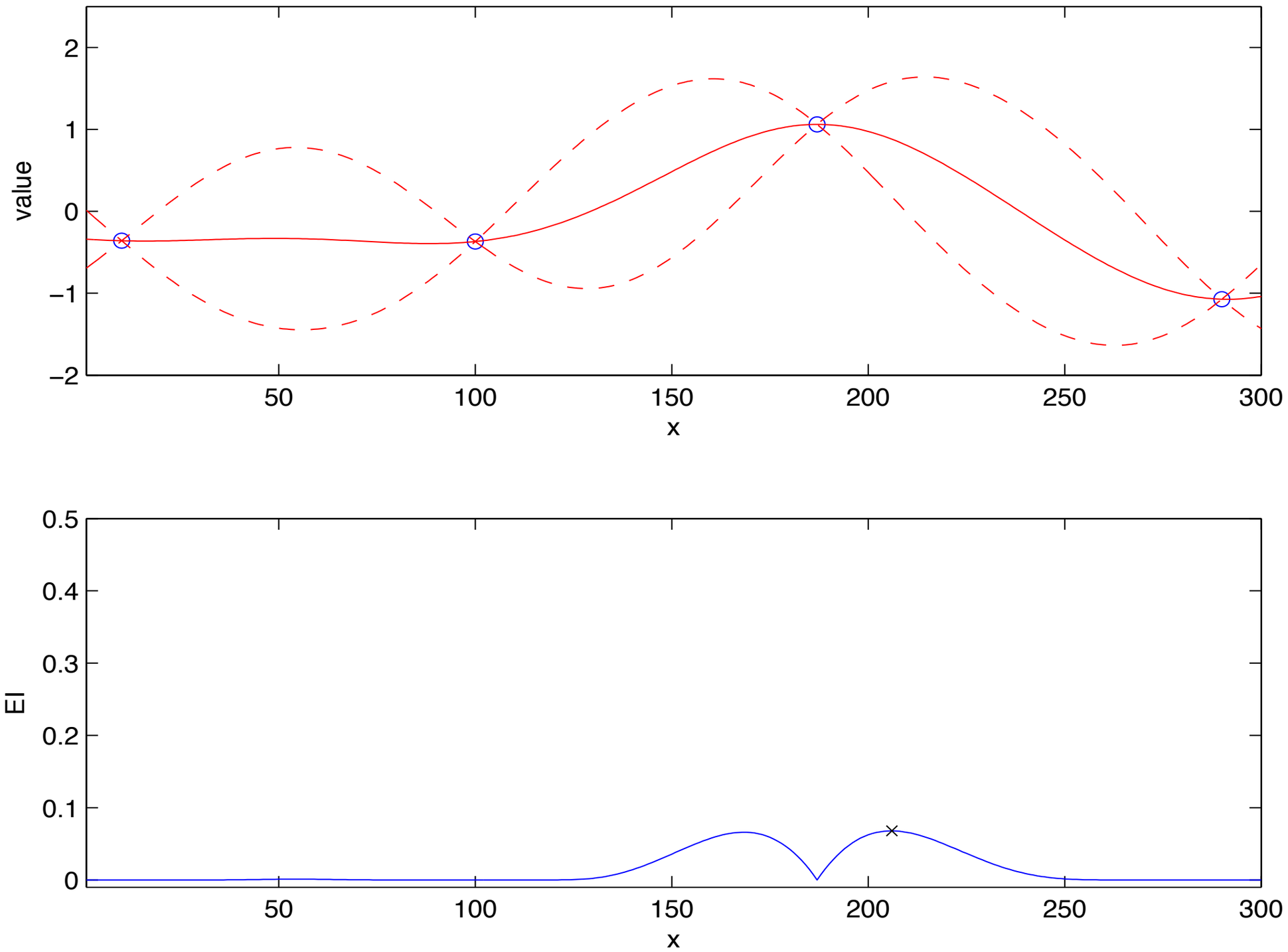
$$\max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \int F\left(x, w\right) p(w) dw$$

- F is **continuous** on x.
- F is **derivative-free**, and **expensive** to evaluate.
- F may be **noisy**.

Background: Gaussian Process Regression



Background: Expected Improvement



Standard Bayesian Optimization is VERY inefficient in this problem

- Model the integral G with a Gaussian process prior.
- while (budget is not exhausted):
  - Find x that maximizes acquisition(x, posterior).
  - Sample x and observe G(x) (**VERY EXPENSIVE!!!**).
  - Update the posterior distribution on G.

Recall:

$$\max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \sum_{w=1}^{n} F(x, w)\, p(w) \qquad \max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \int F(x, w)\, p(w) dw$$

We model F instead of G

- Model the **INTEGRAND** F with a Gaussian process prior.
- G is a linear function of F

$$\max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \sum_{w=1}^{n} F(x, w)\, p(w)$$

$$\max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \int F(x, w)\, p(w) dw$$

- *Bayesian Quadrature* implies a Gaussian process prior on G.

The challenge is choosing where to sample.

Previous Bayesian Optimization Approaches

- Williams et al. 2000: Problem (1) when F is noiseless. It fails to be consistent. It “Hacks” EI.

- Groot et al. 2010: Problem (2) when F is noiseless, p and kernel are both Gaussian. There is no improvement over evaluating *G* directly.

- Xie et al. 2012: Problem (2) when *F* is noiseless, p and kernel are both Gaussian. Infeasible when the dimension of domain is bigger than 3.

- Swersky et al. 2013: Problem (1) when F is noiseless. Infeasible when there are many summands. It “Hacks” EI.

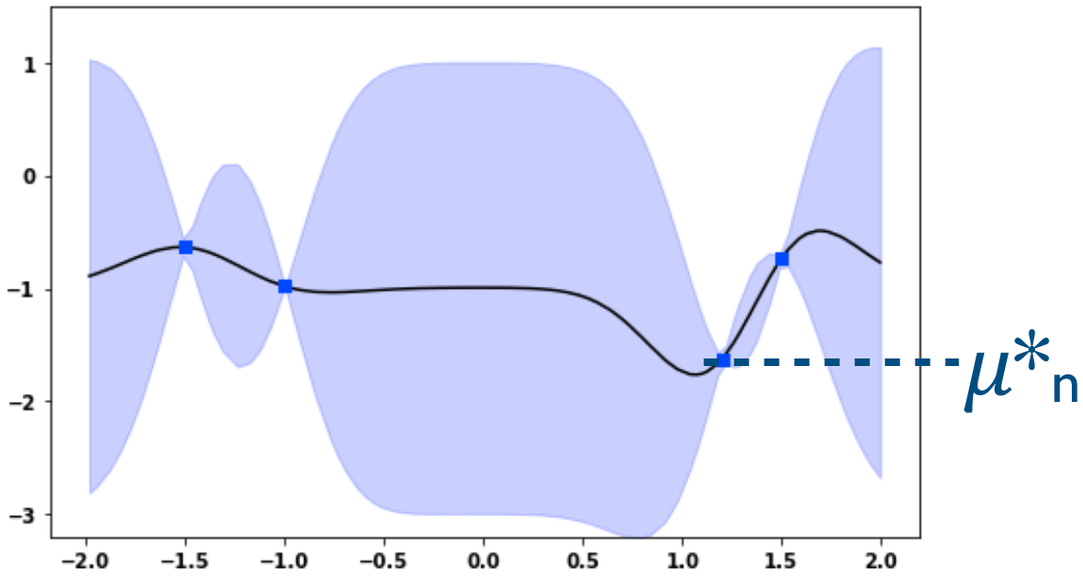
(1) $\max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \sum_{w=1}^{n} F(x, w)\, p(w)$ **(2)** $\max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \int F(x, w)\, p(w) dw$

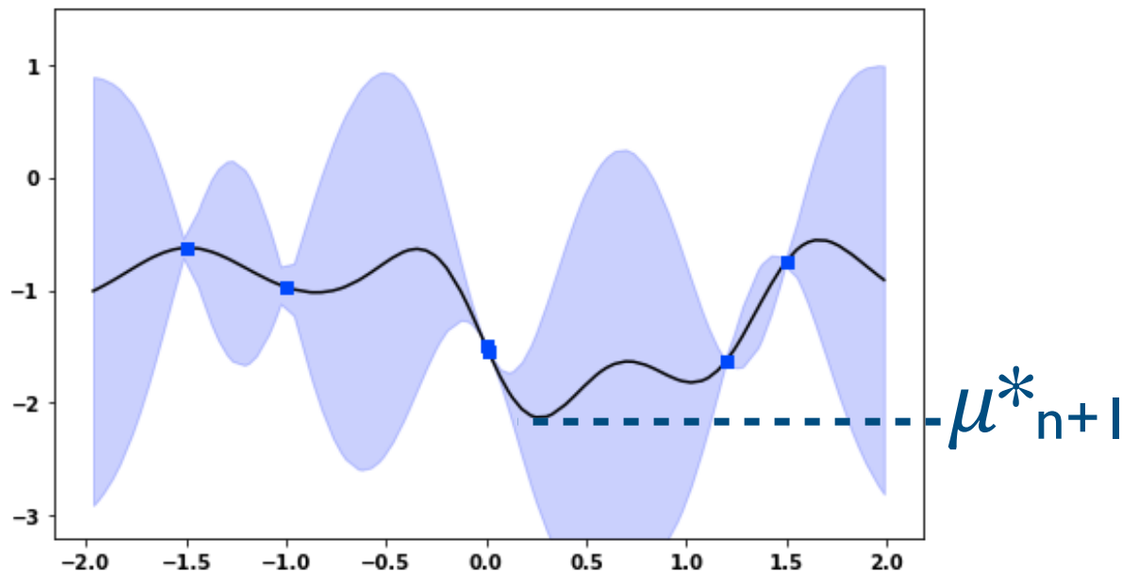
Our contribution

We provide a method for choosing where to sample, Bayesian Quadrature Optimization (BQO), with these advantages:

- It is **more general** (sums or integrals, no restrictions on p or kernel, allows noise).
- It has more substantial **theoretical justification** (one-step optimal; asymptotically consistent).
- It has better **empirical performance**.

Bayesian Quadrature Optimization





- Our acquisition function is based on choosing the optimal point when we have only one more sample to take (*knowledge gradien*t (KG) technique).

- Reduction in loss due to sampling:
  $BQO(x,w) = E_n[\mu^*_n - \mu^*_{n+1} \mid \text{sample } x,w]$

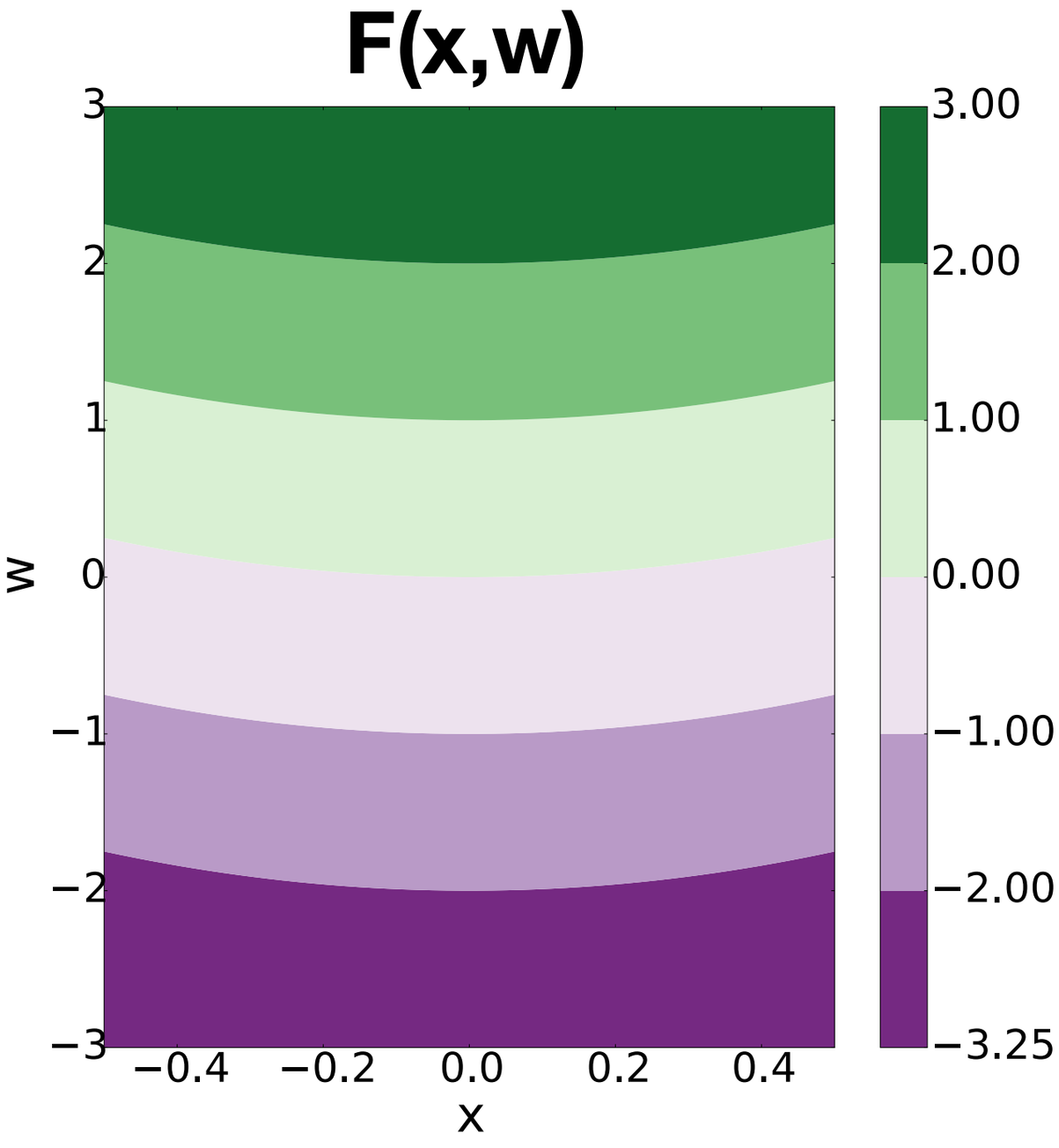
$\mu_n(x) := E_n[G(x)]$ is the expected value of our objective under the posterior @ time n

We develop a novel and efficient discretization-free computational method for optimizing BQO

- Estimate ∇BQO(x, w) using infinitesimal perturbation analysis (IPA) and the envelope theorem.

- Use multistart stochastic gradient ascent to find an approximate solution to solve argmax BQO(x, w).

Here's an example



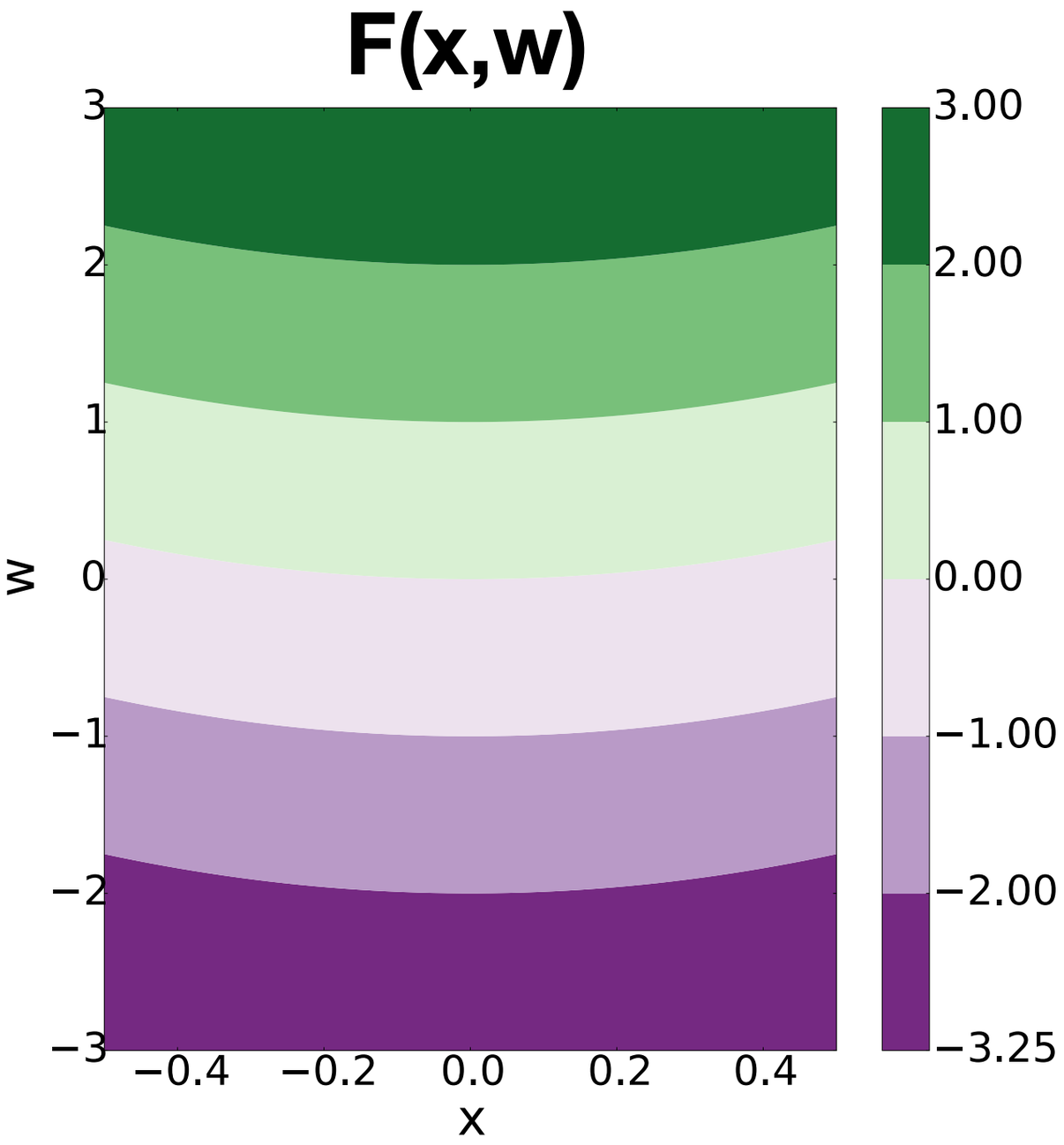
$F(x,w) = E[zx^2 + w \mid w]$,
where $w \sim N(0,1)$ and $z \sim N(-1,1)$

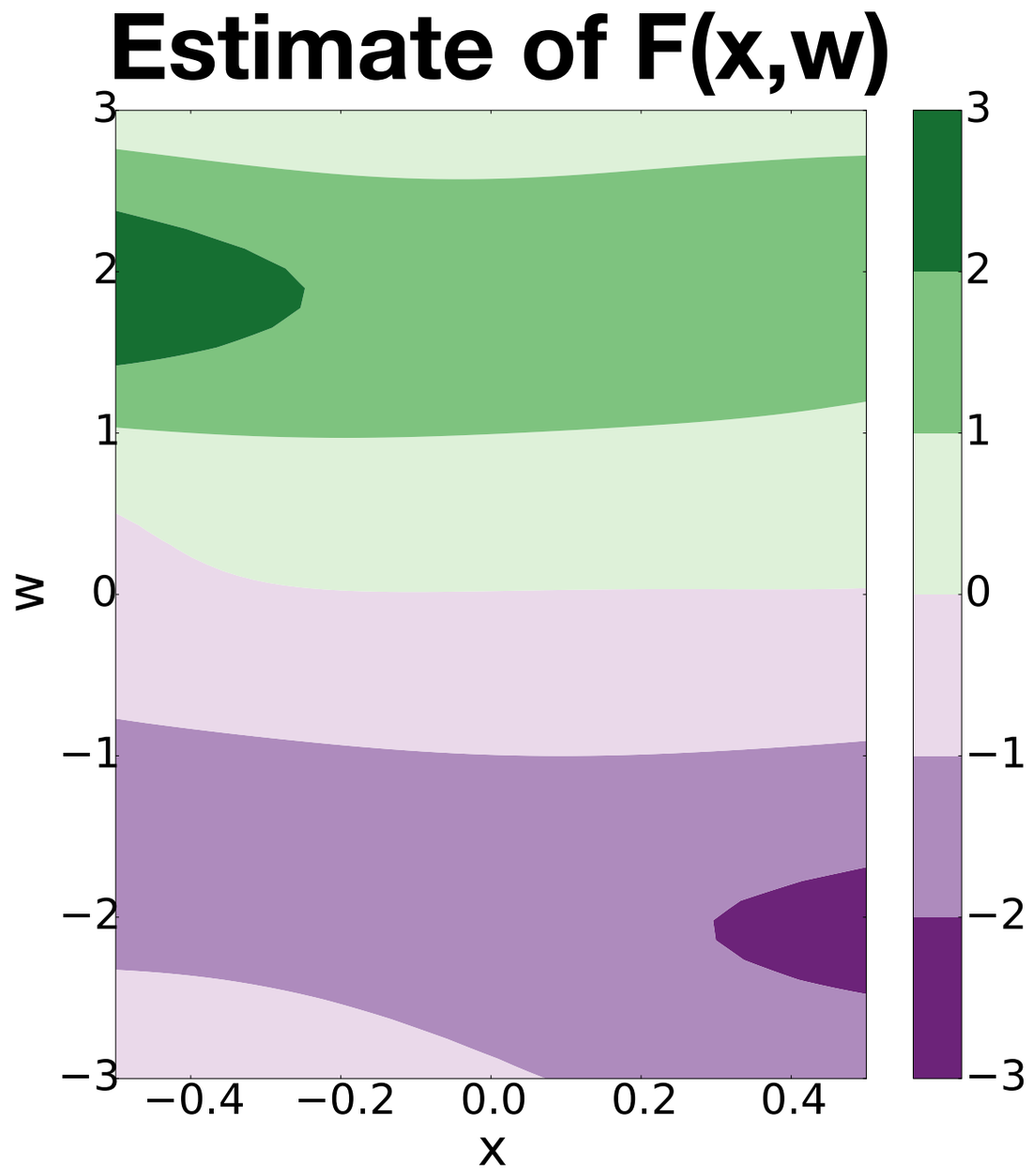
To sample F(x,w),
we sample $zx^2 + w$

$G(x) = E[zx^2 + w] = E[F(x,w)]$

We want to solve:
max_x G(x)

Here's BQO's estimate of F after a few samples

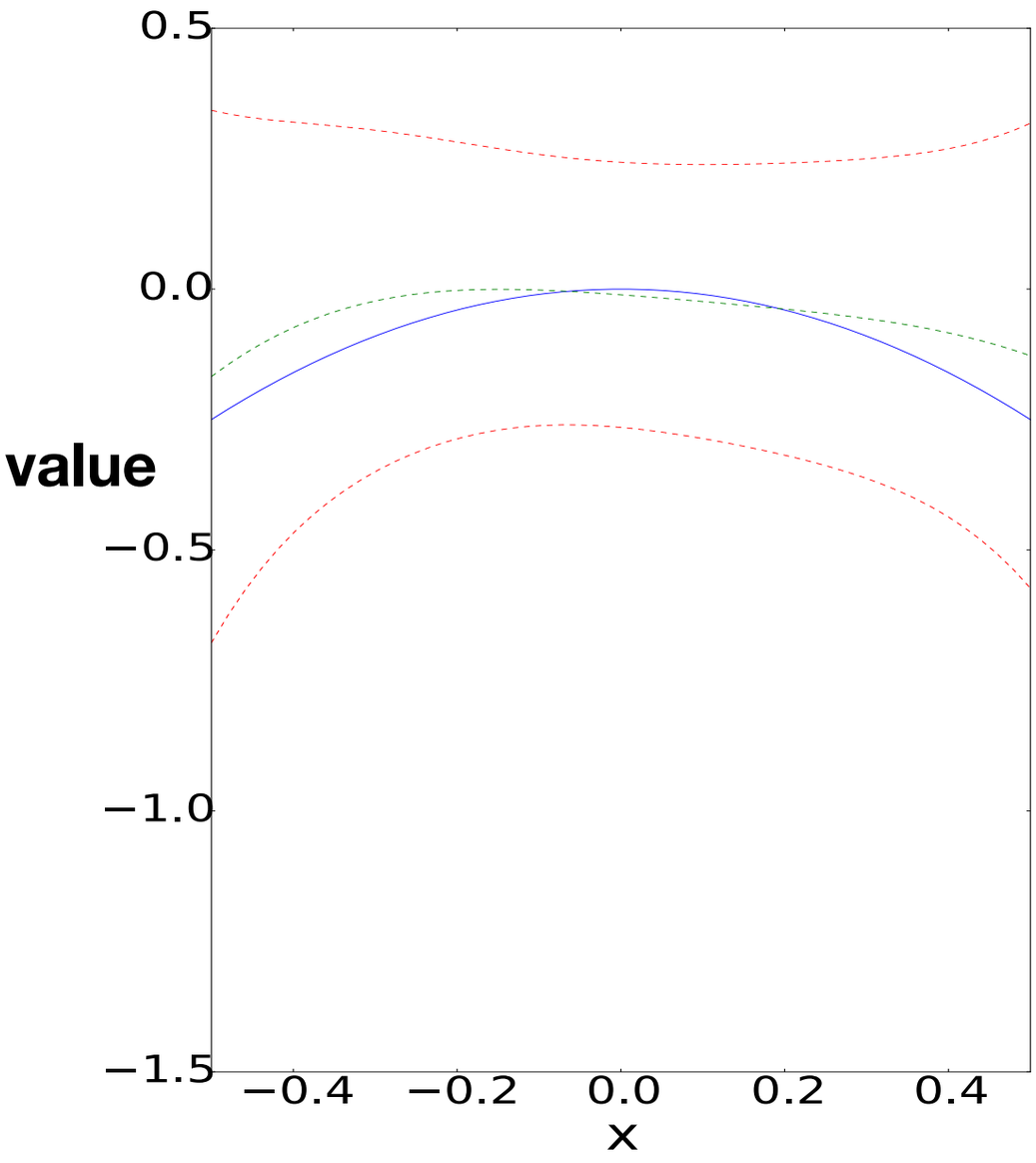




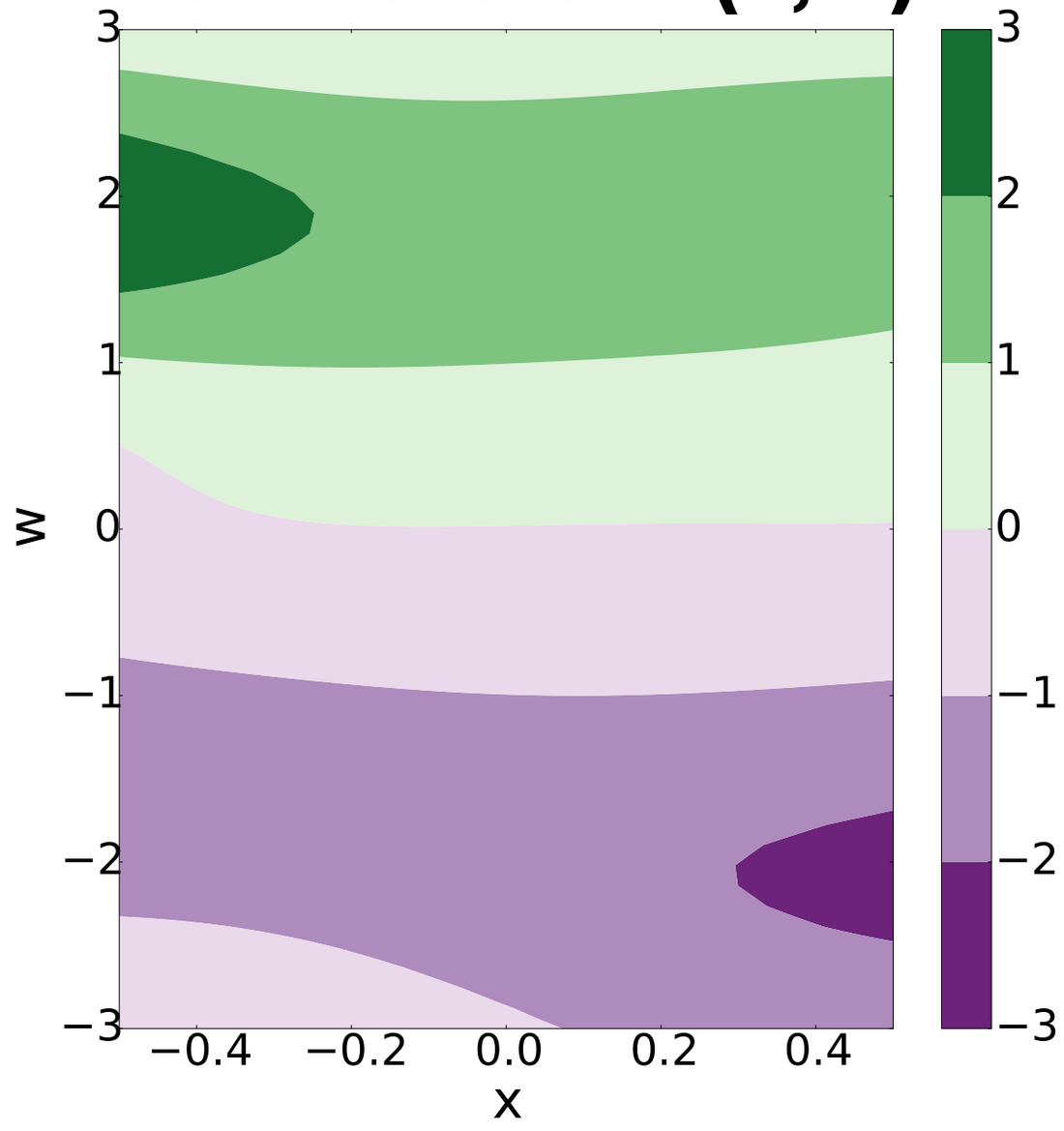
Recall: $F(x,w) = E[zx^2 + w \mid w]$,
$w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

Here's BQO's estimate of G(x) after a few samples

G(x) & BQO's estimate + CI

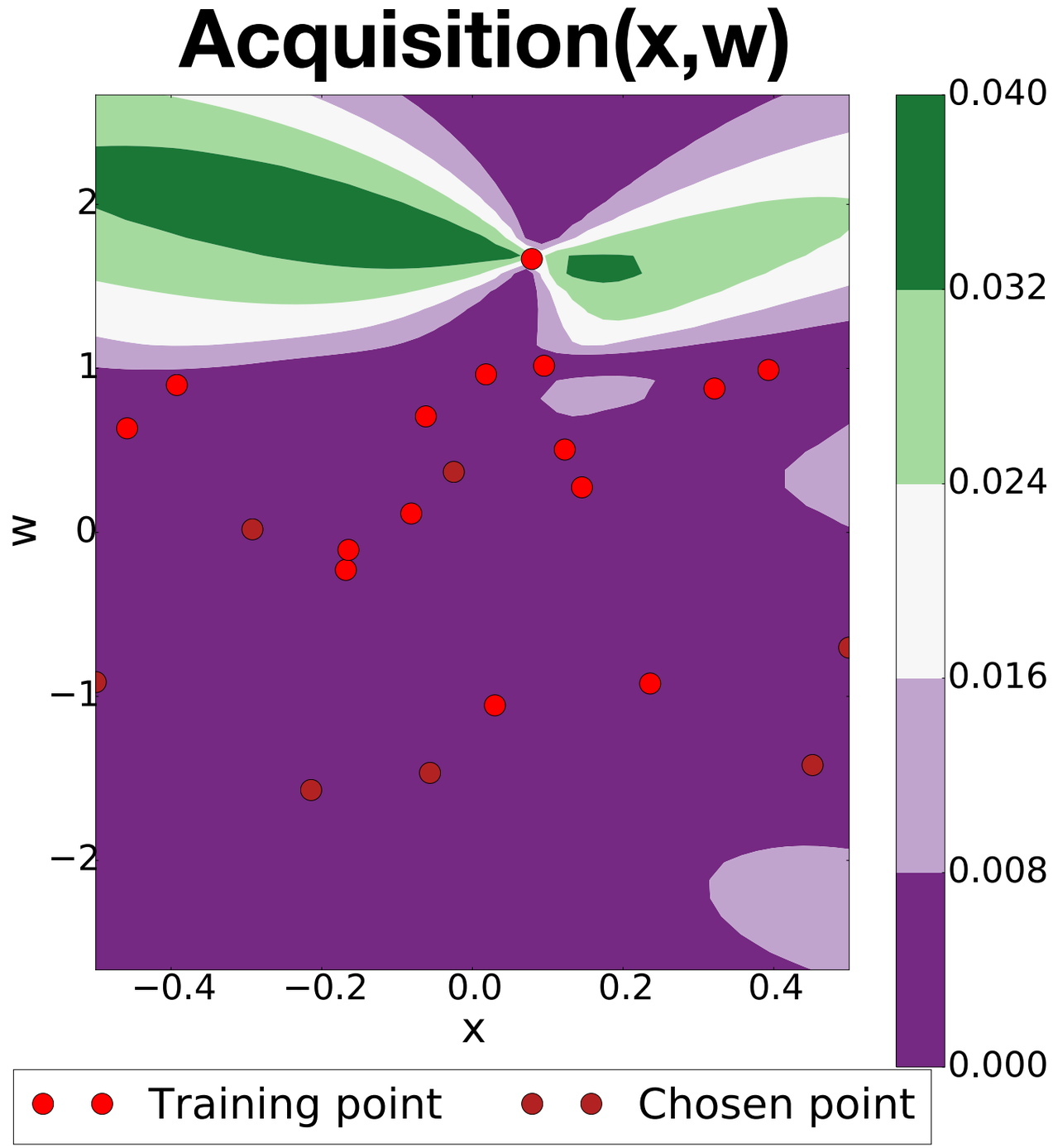


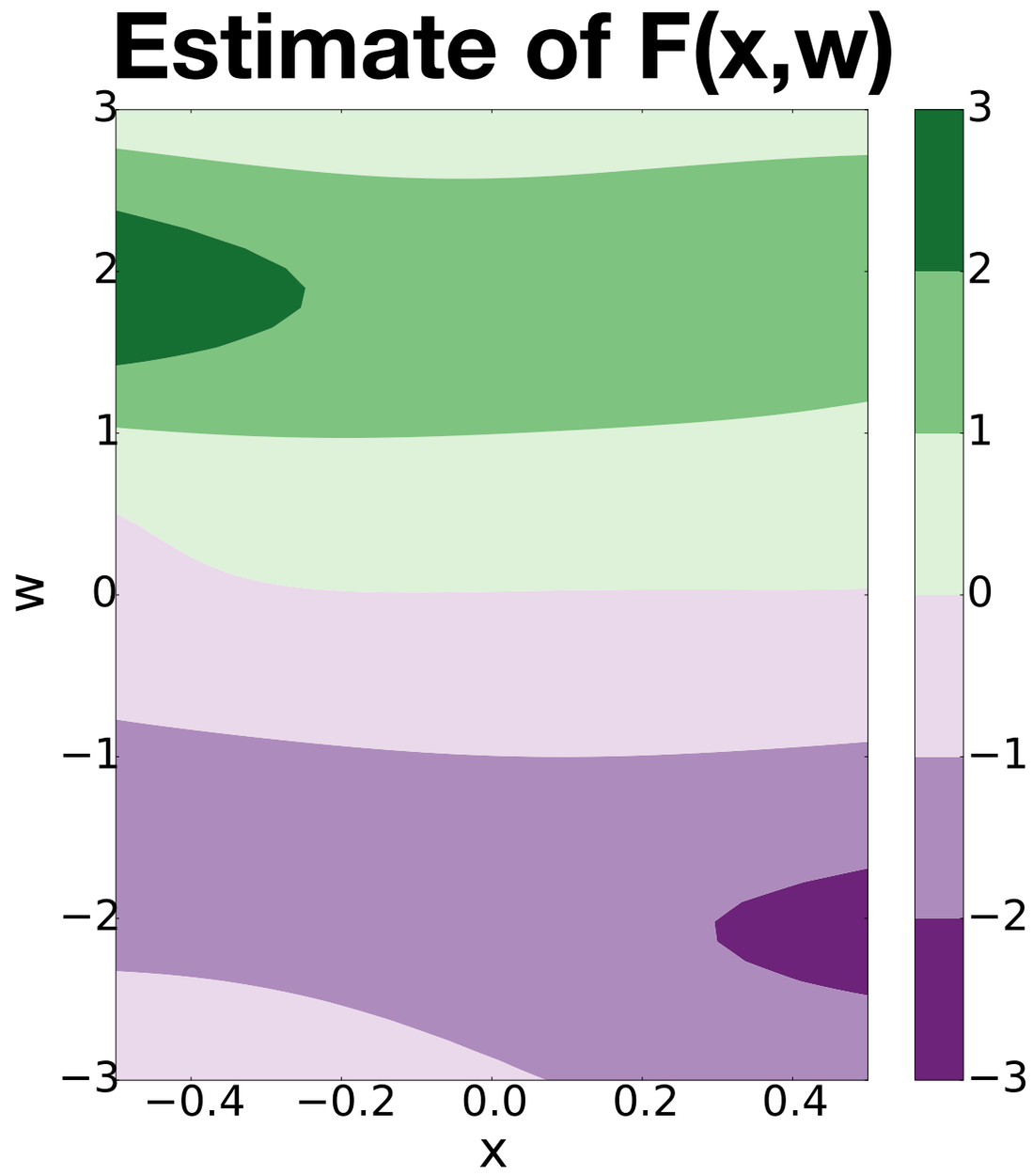
Estimate of F(x,w)



Recall: $F(x,w) = E[zx^2 + w \mid w]$,
$w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

Here's BQO's acquisition function after a few samples

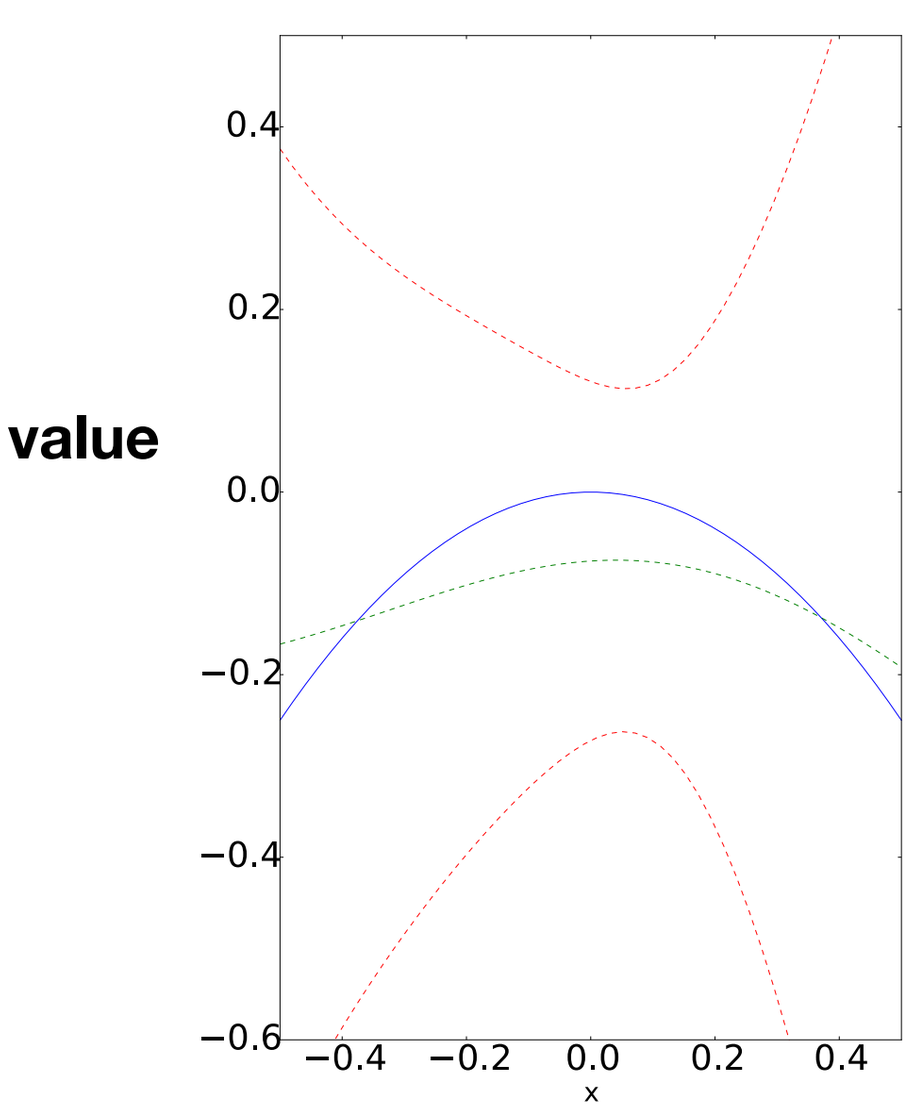




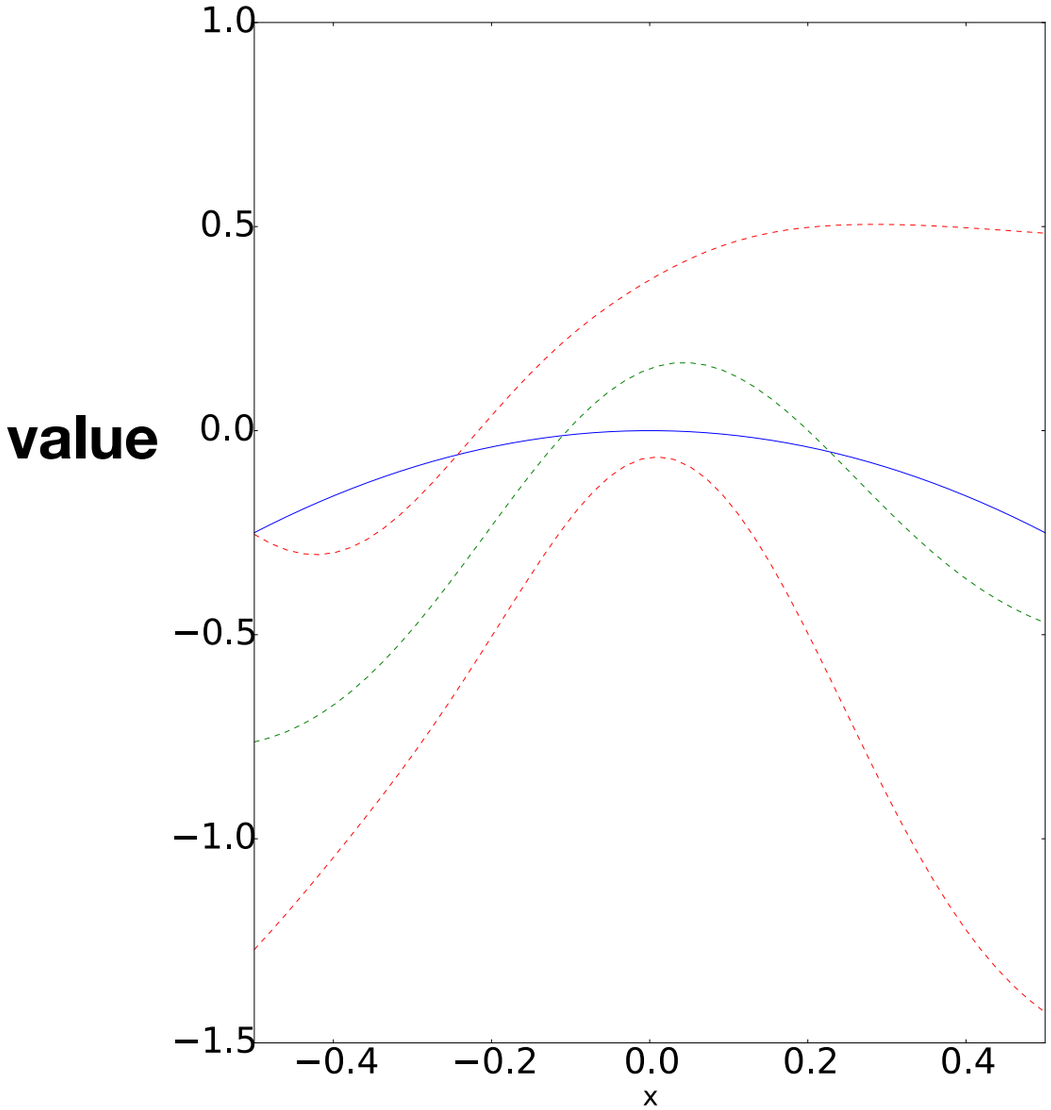
Recall: $F(x,w) = E[zx^2 + w \mid w]$,
$w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

BQO explores more effectively than KG

G(x) & BQO's estimate + CI



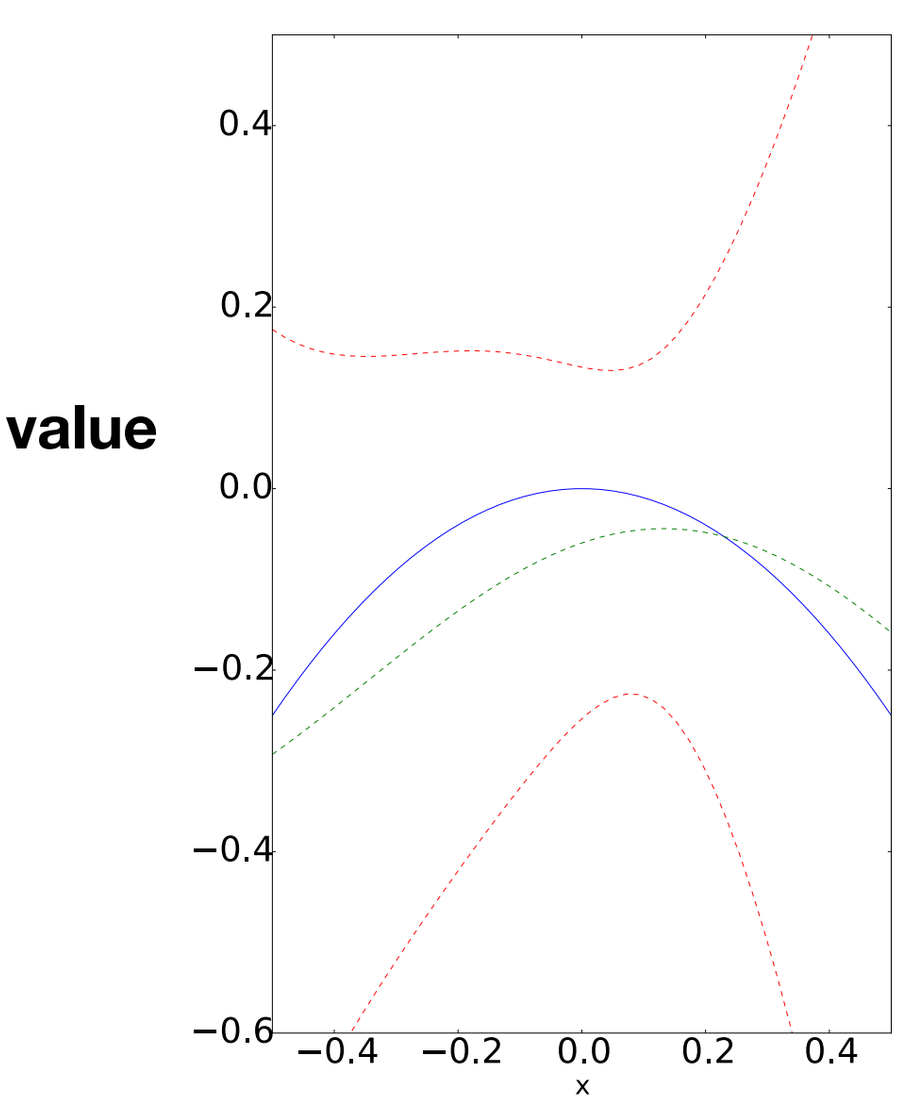
G(x) & KG's estimate + CI



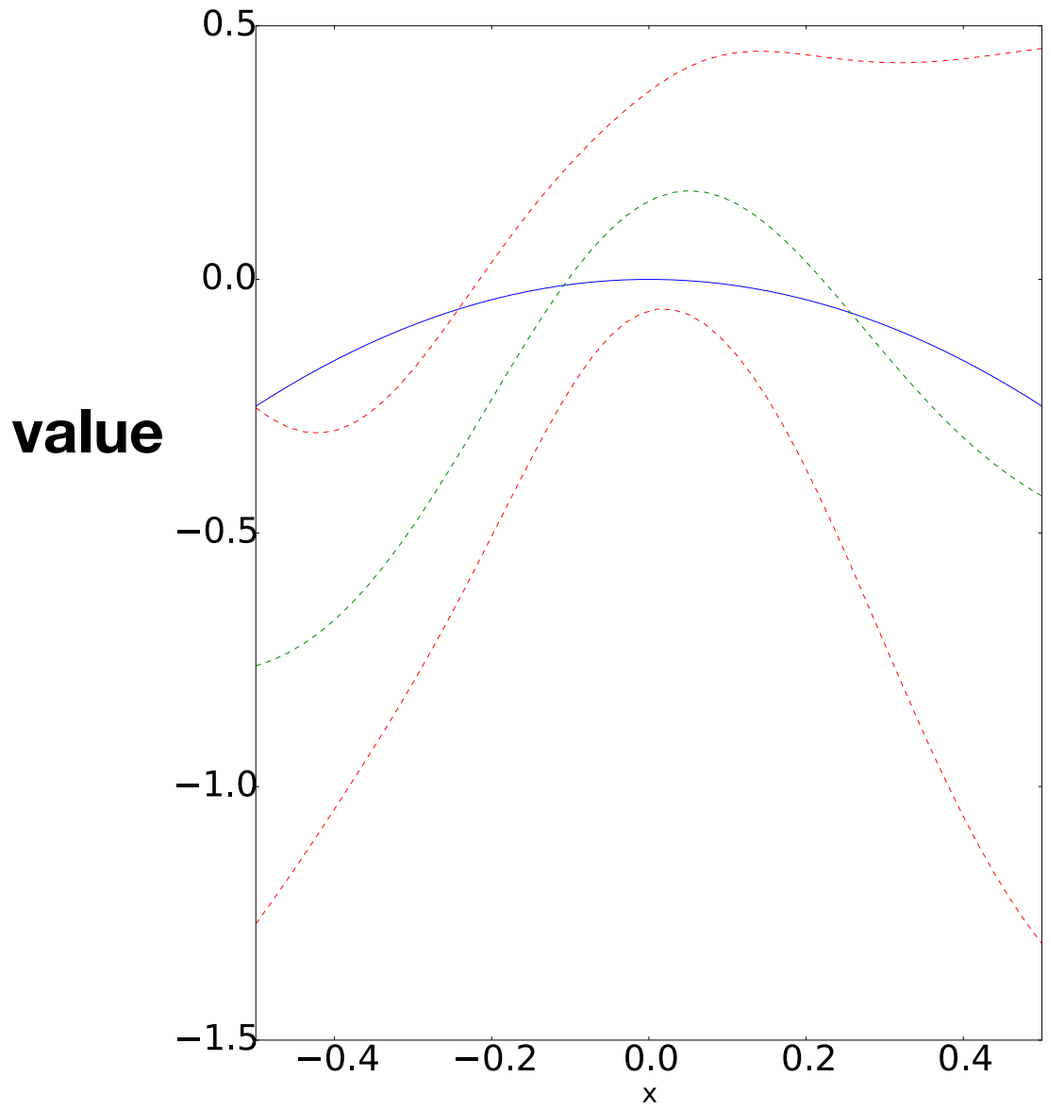
Recall: $F(x,w) = E[zx^2 + w \mid w]$,
$w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

BQO explores more effectively than KG

G(x) & BQO's estimate + CI



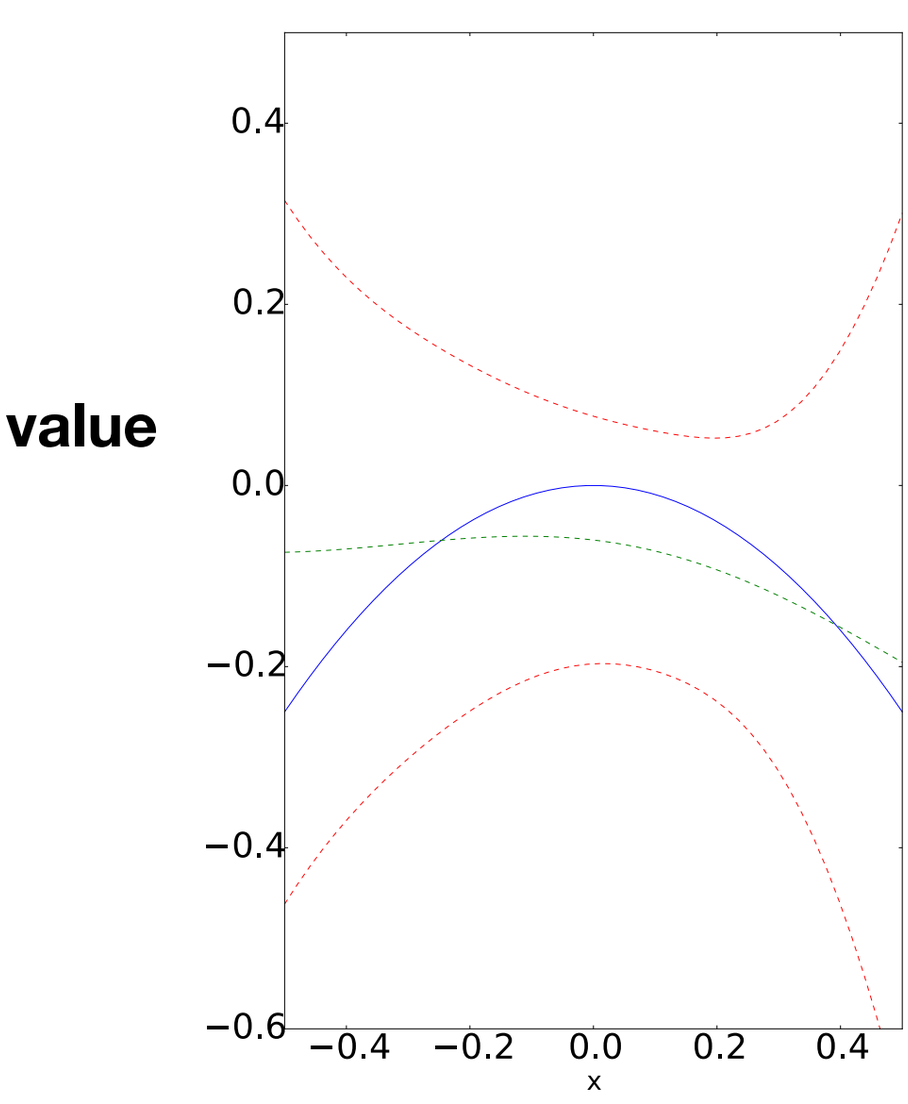
G(x) & KG's estimate + CI



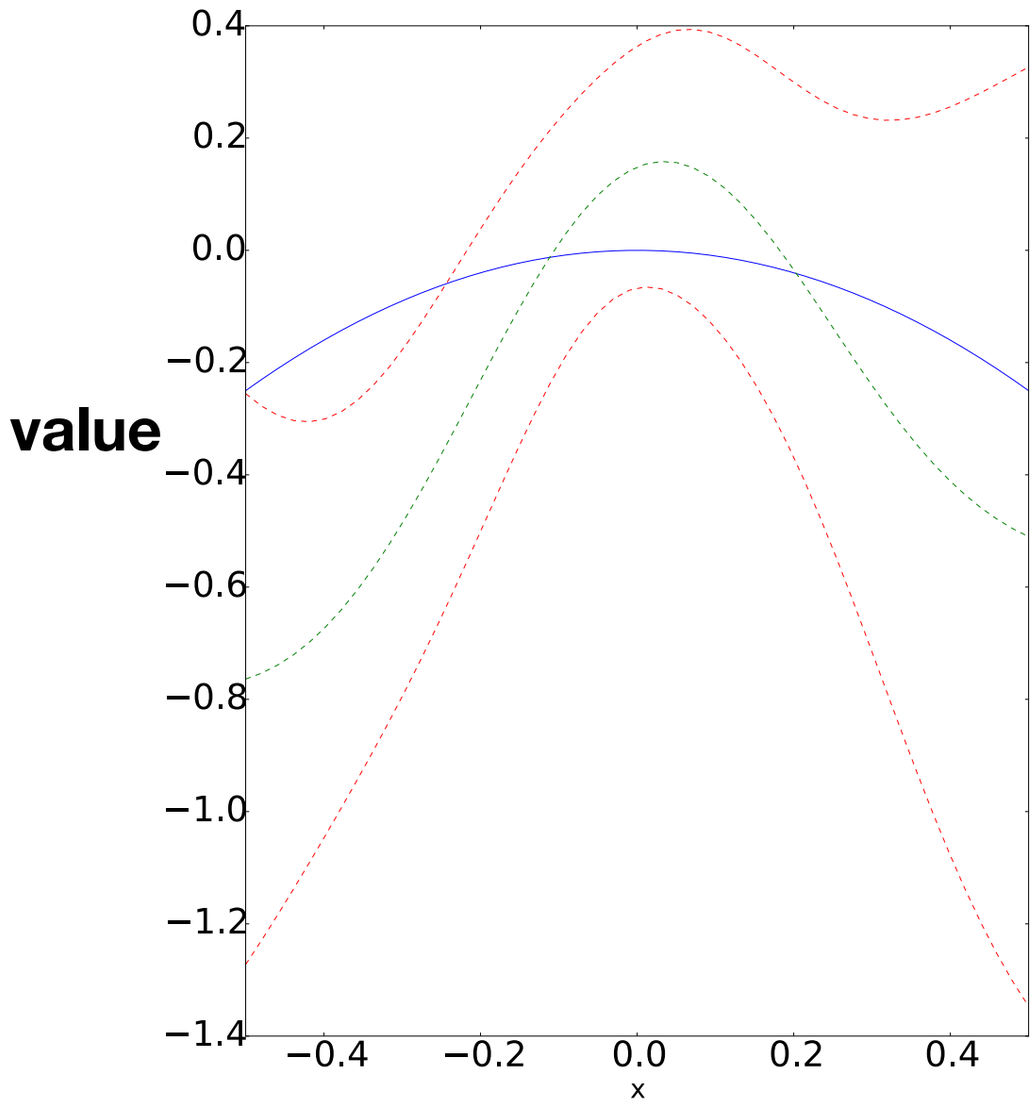
Recall: $F(x,w) = E[zx^2 + w \mid w]$, $w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

BQO explores more effectively than KG

G(x) & BQO's estimate + CI



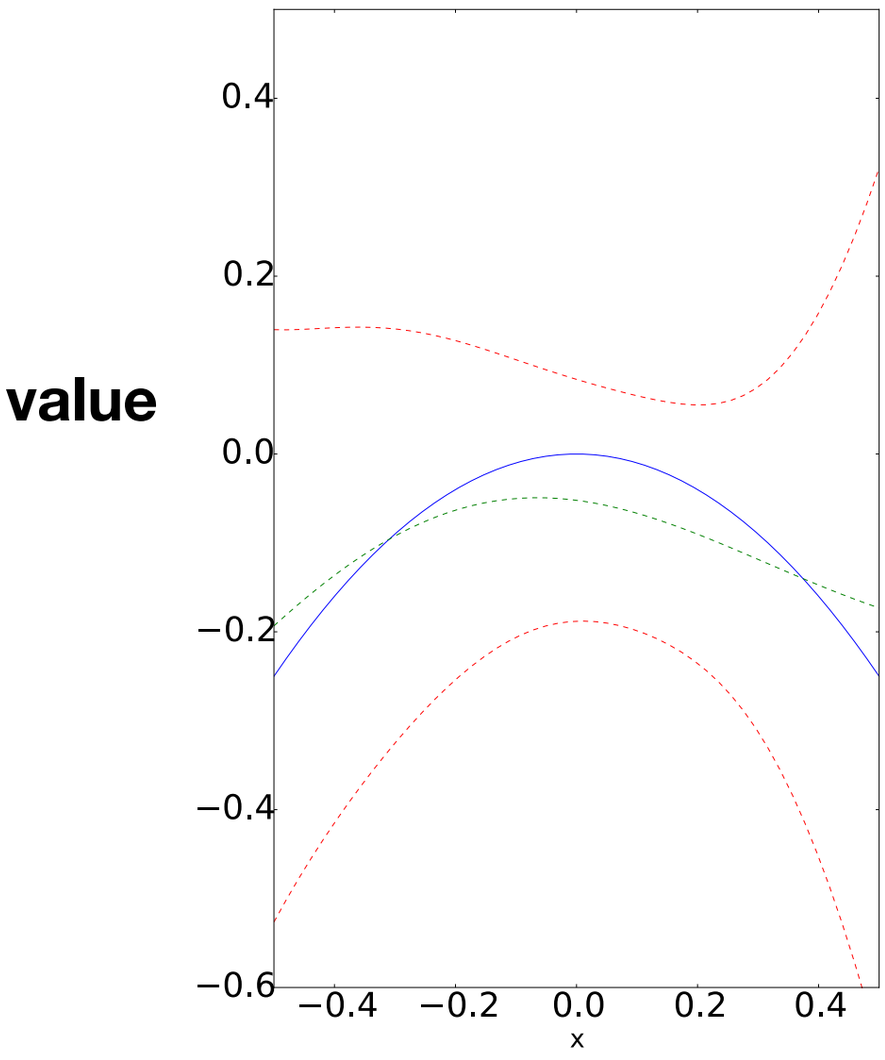
G(x) & KG's estimate + CI



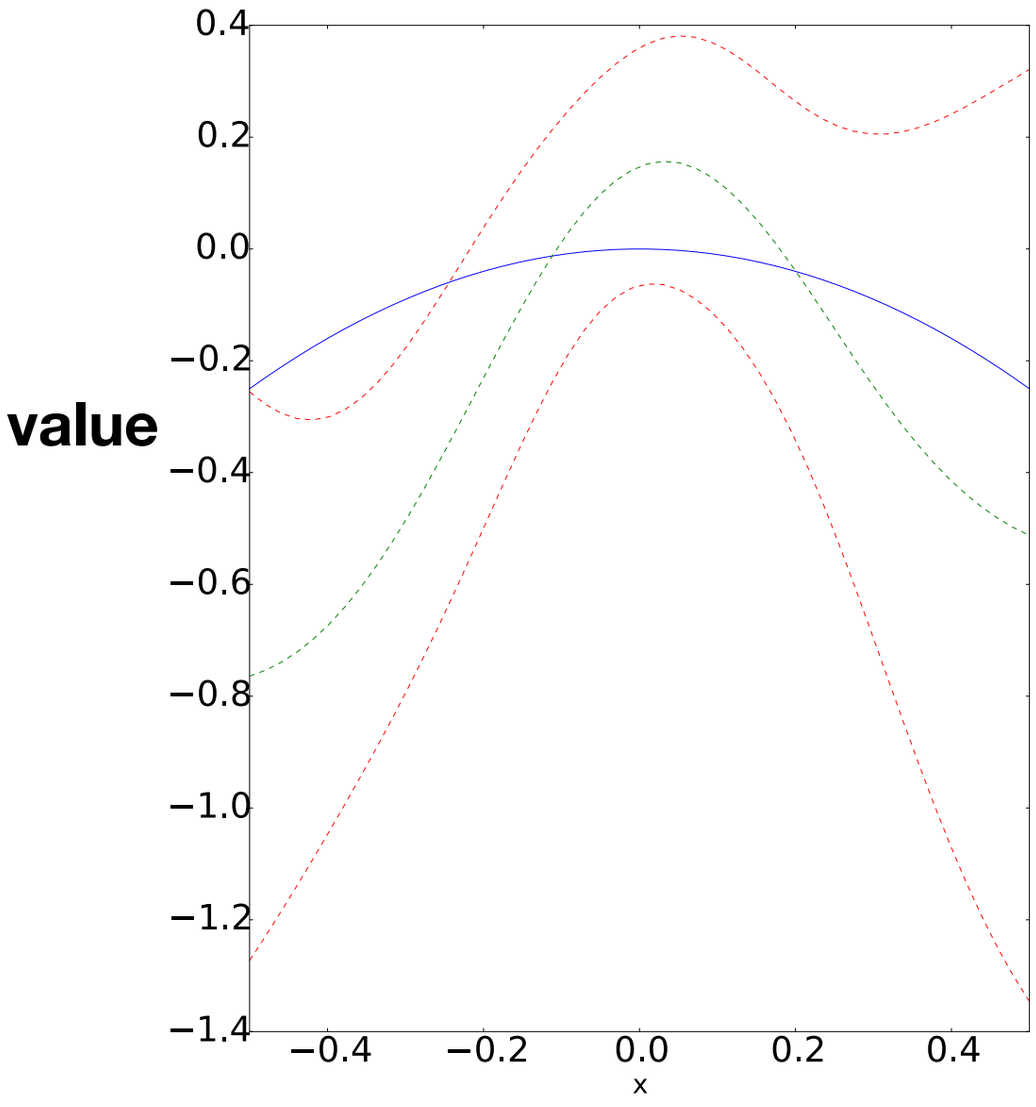
Recall: $F(x,w) = E[zx^2 + w \mid w]$, $w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

BQO explores more effectively than KG

G(x) & BQO's estimate + CI



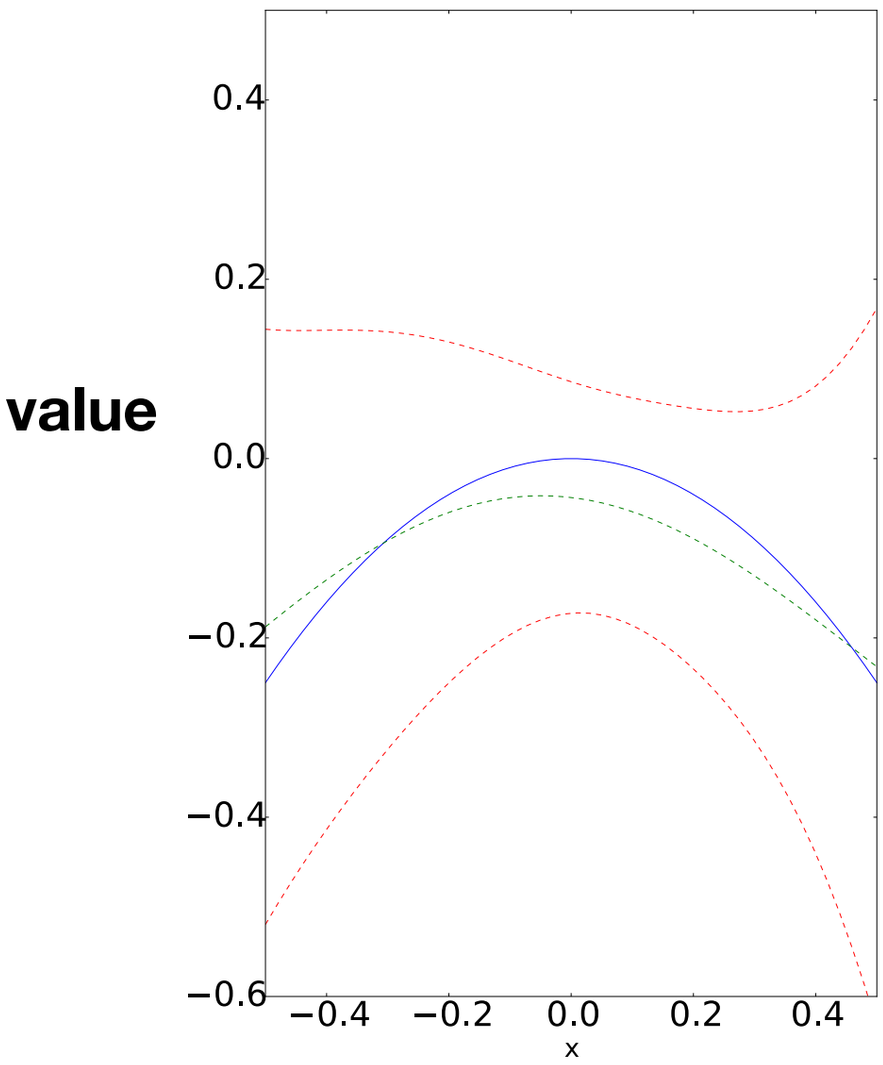
G(x) & KG's estimate + CI



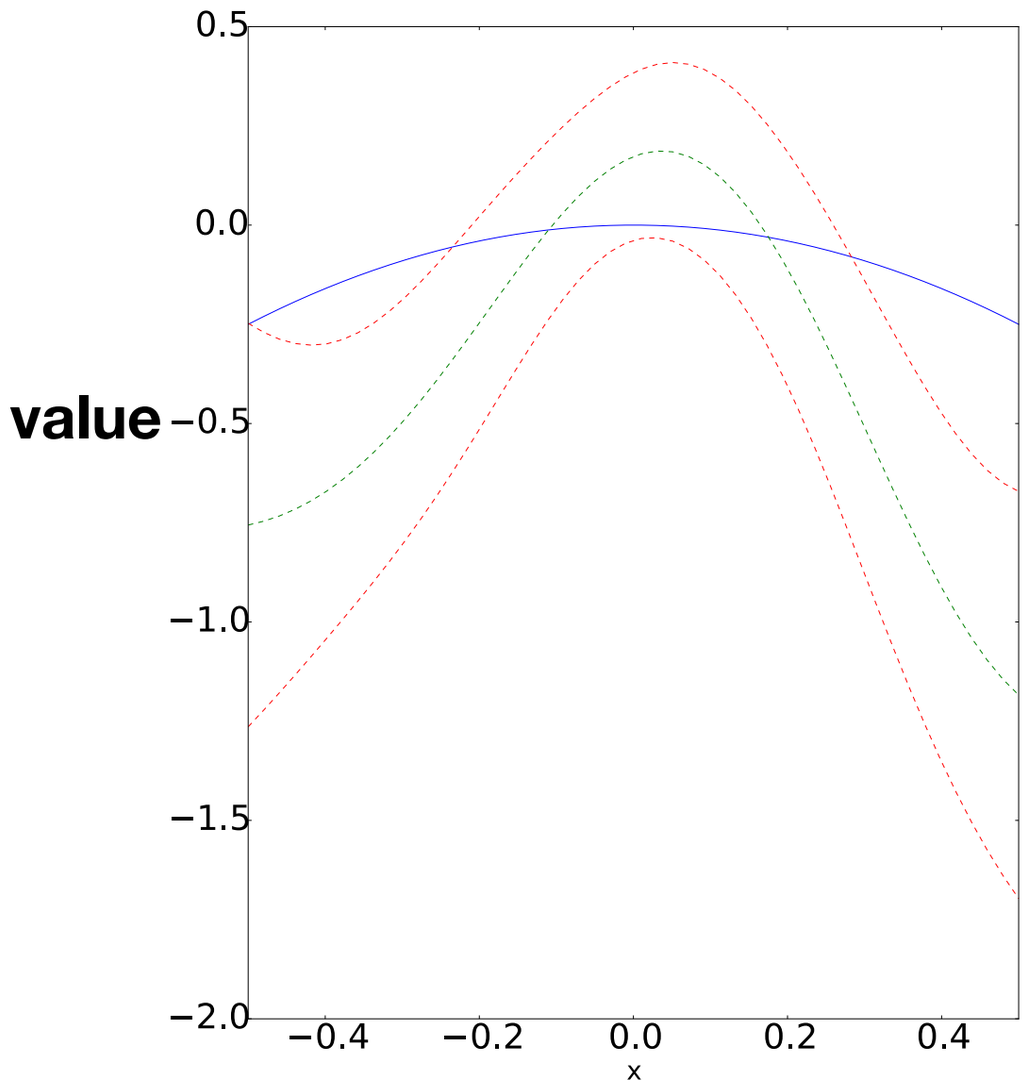
Recall: $F(x,w) = E[zx^2 + w \mid w]$, $w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

BQO explores more effectively than KG

G(x) & BQO's estimate + CI



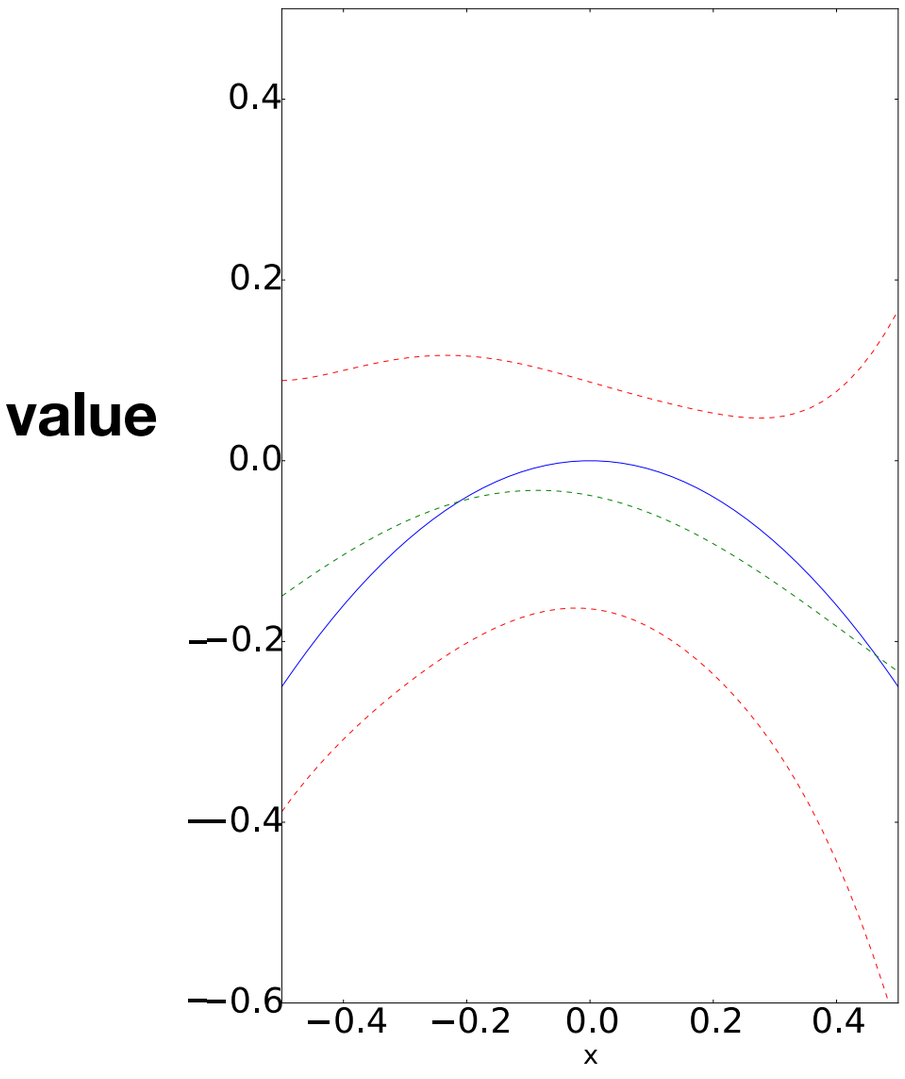
G(x) & KG's estimate + CI



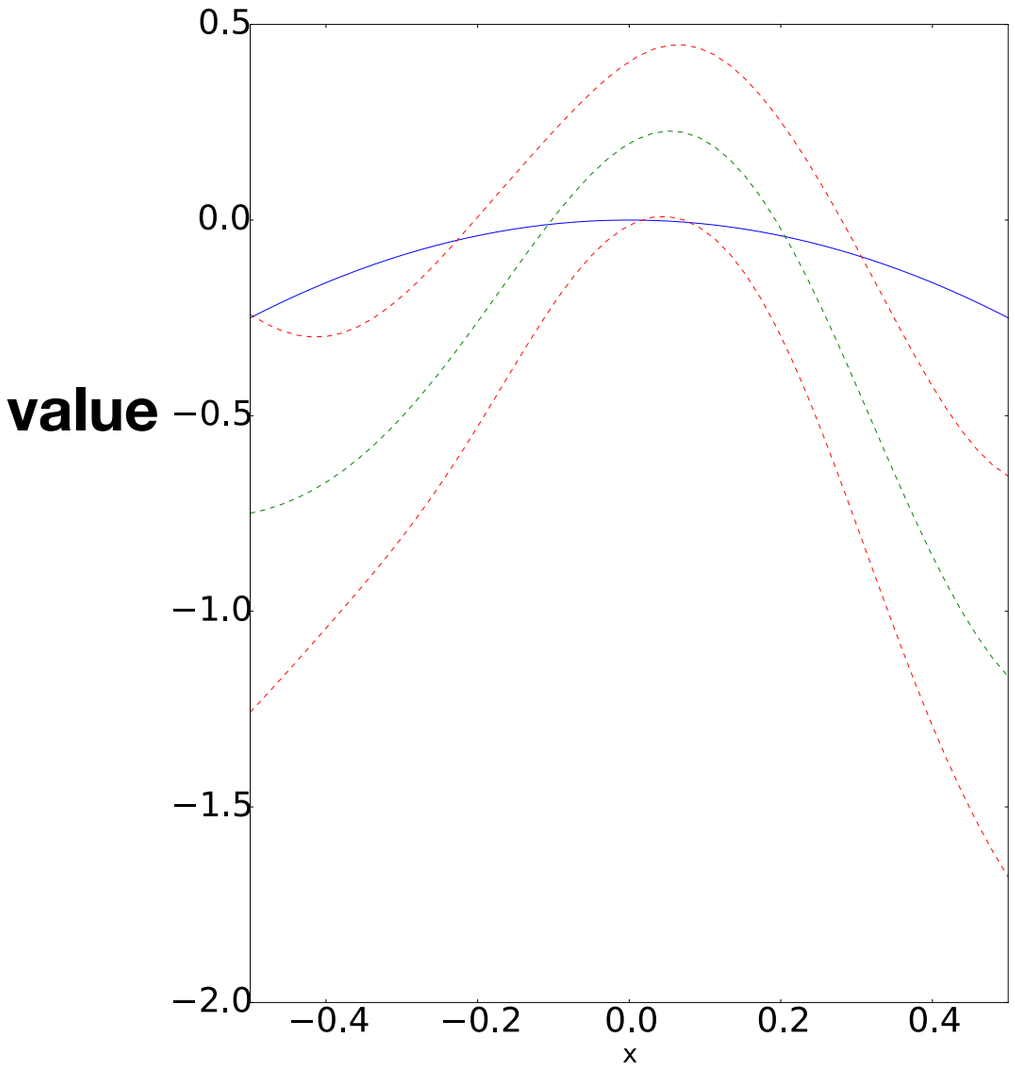
Recall: $F(x,w) = E[zx^2 + w \mid w]$,
$w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

BQO explores more effectively than KG

G(x) & BQO's estimate + CI



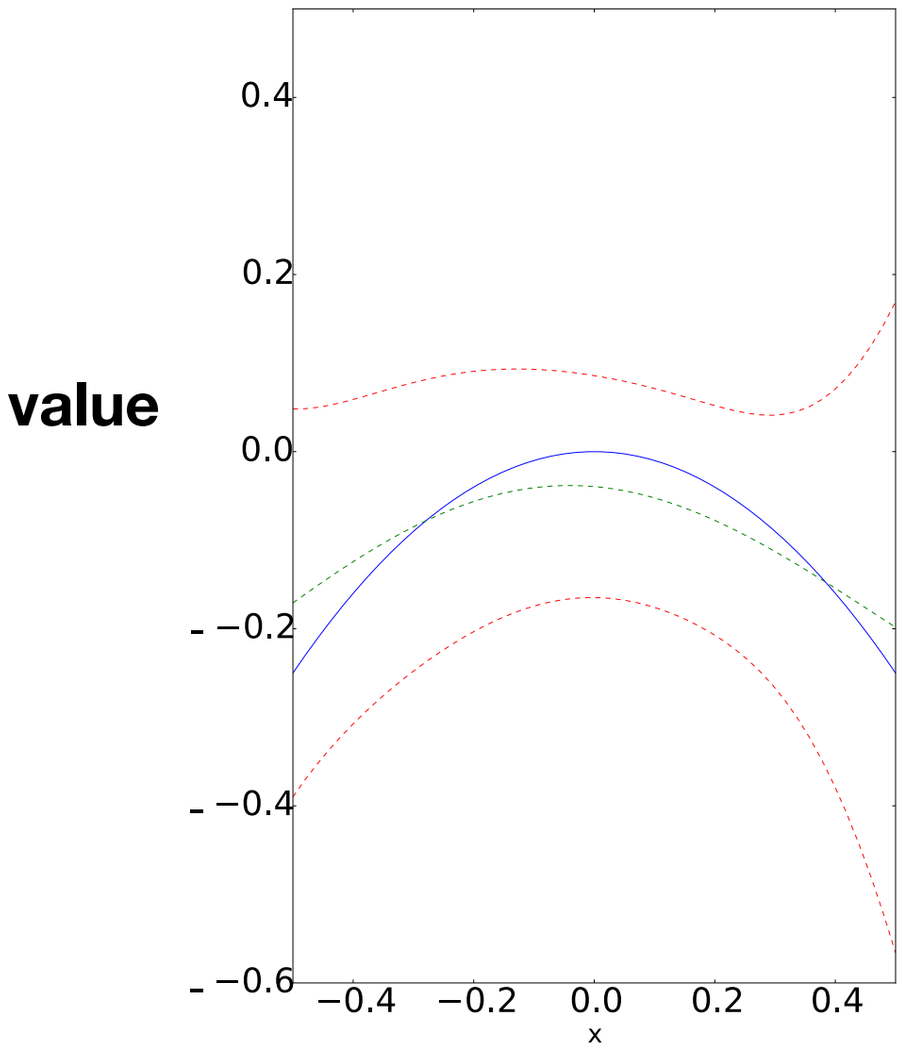
G(x) & KG's estimate + CI



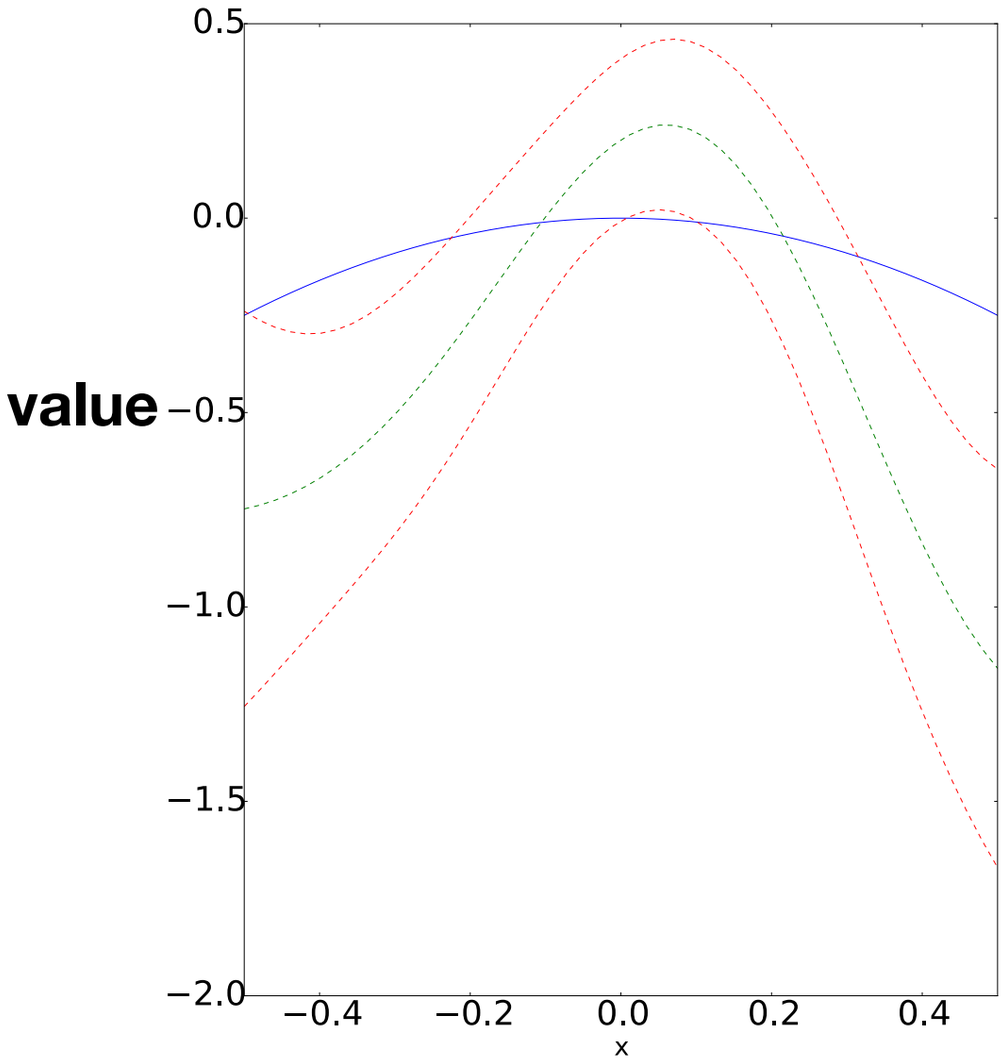
Recall: $F(x,w) = E[zx^2 + w \mid w]$,
$w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

BQO explores more effectively than KG

G(x) & BQO's estimate + CI



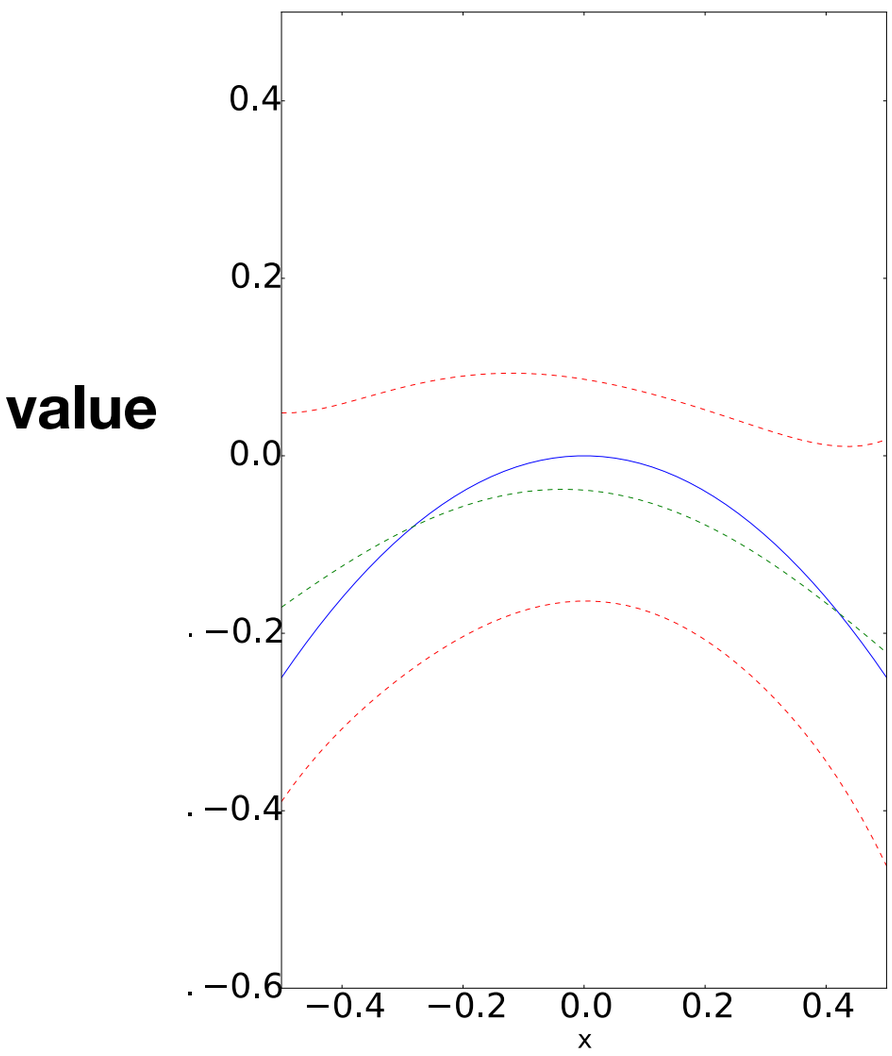
G(x) & KG's estimate + CI



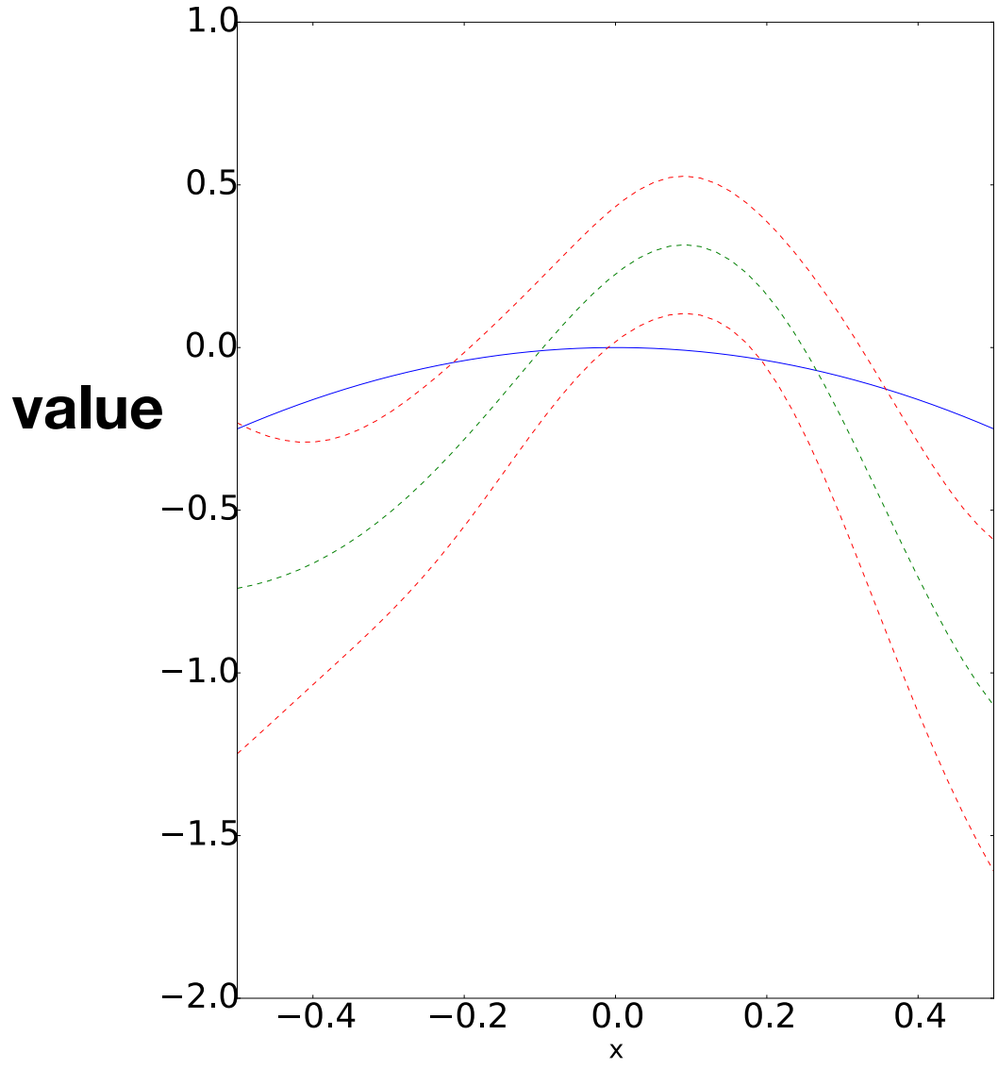
Recall: $F(x,w) = E[zx^2 + w \mid w]$, $w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

BQO explores more effectively than KG

G(x) & BQO's estimate + CI



G(x) & KG's estimate + CI



Recall: $F(x,w) = E[zx^2 + w \mid w]$,
$w \sim N(0,1)$ and $z \sim N(-1,1)$
$G(x) = E[zx^2 + w] = E[F(x,w)]$

Theoretical Results

Theorem 1. If the discretization-free computation technique is used to estimate stochastic gradients of BQO, then those estimators are unbiased and strongly consistent.

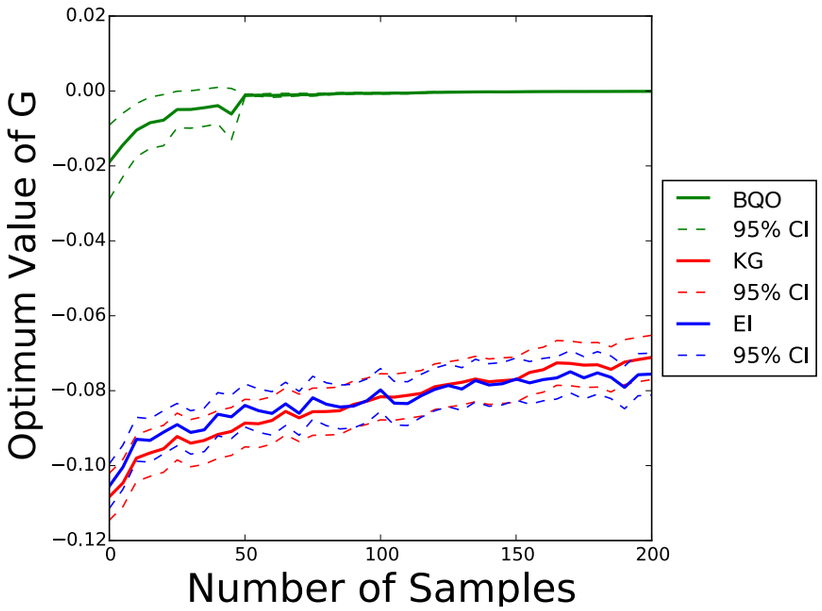
Theorem 2. If the domain of x is finite or continuous in the problem of the sum, then BQO finds an optimal solution if the number of samples tends to infinity.

Recall:

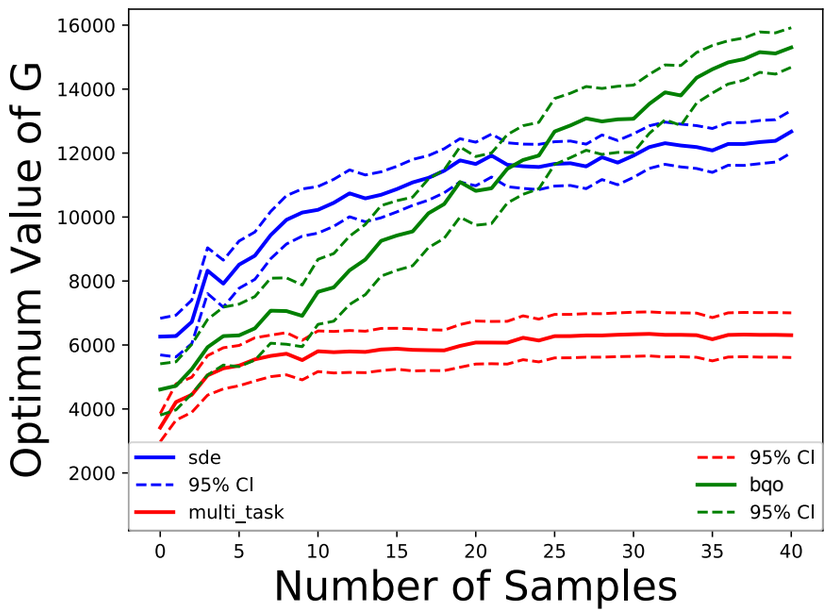
$$\max_{x \in A \subset \mathbb{R}^d} G(x) := \max_{x \in A \subset \mathbb{R}^d} \sum_{w=1}^{n} F\left(x, w\right) p(w)$$

Numerical Results

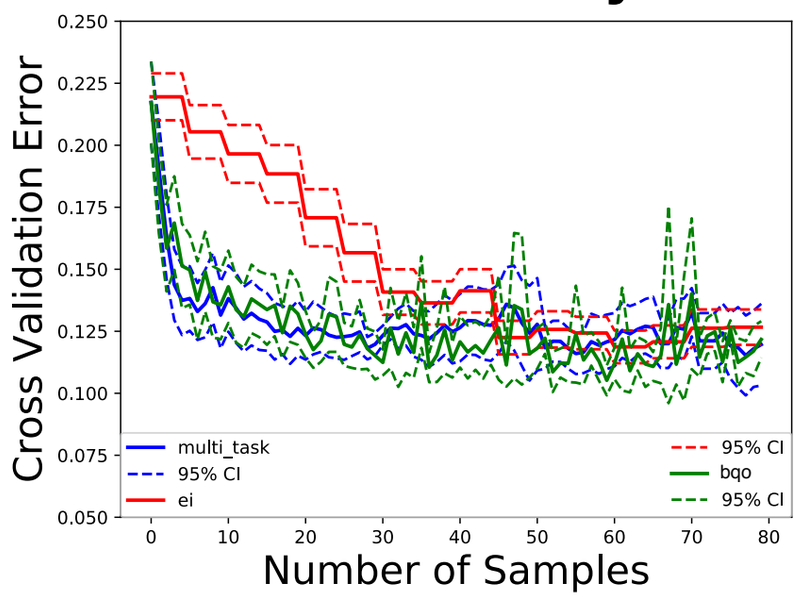
Analytic Example



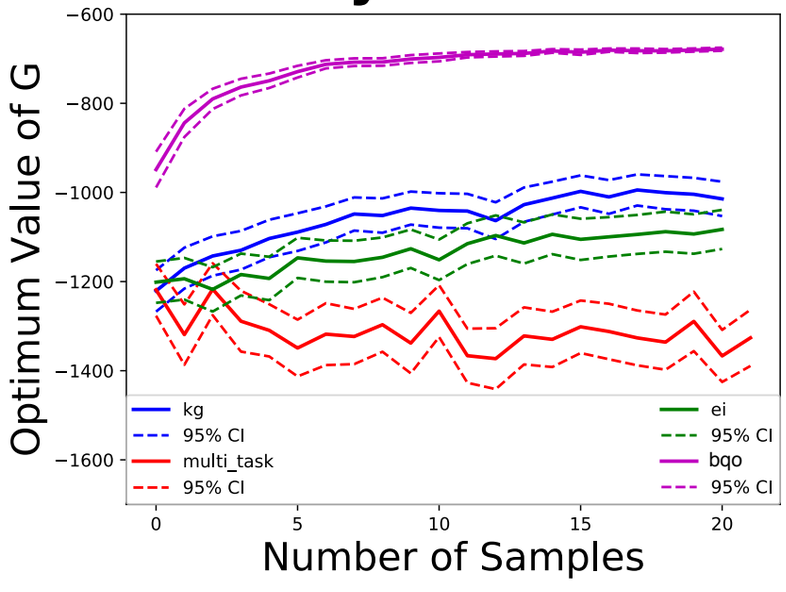
Branin Problem



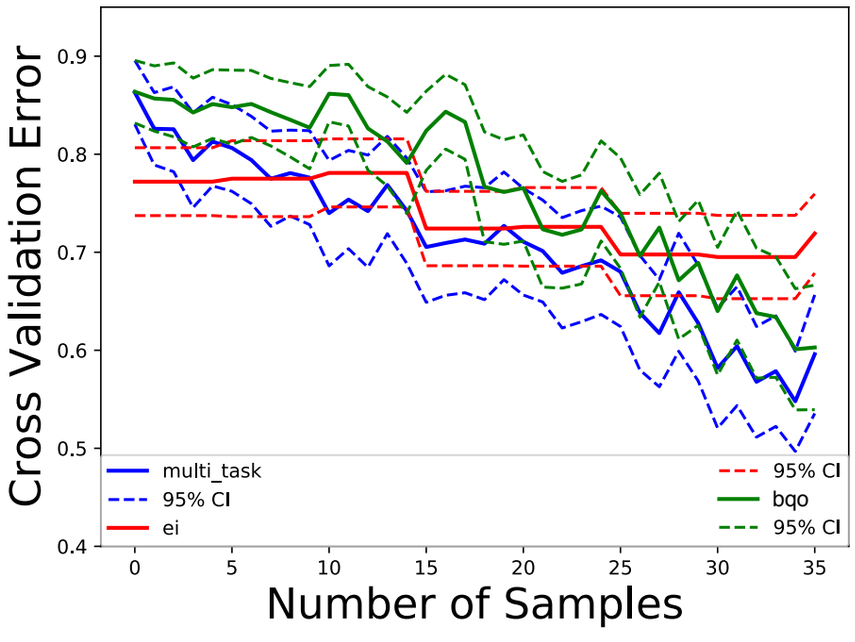
Hyperparameter Tuning in Recommender Systems



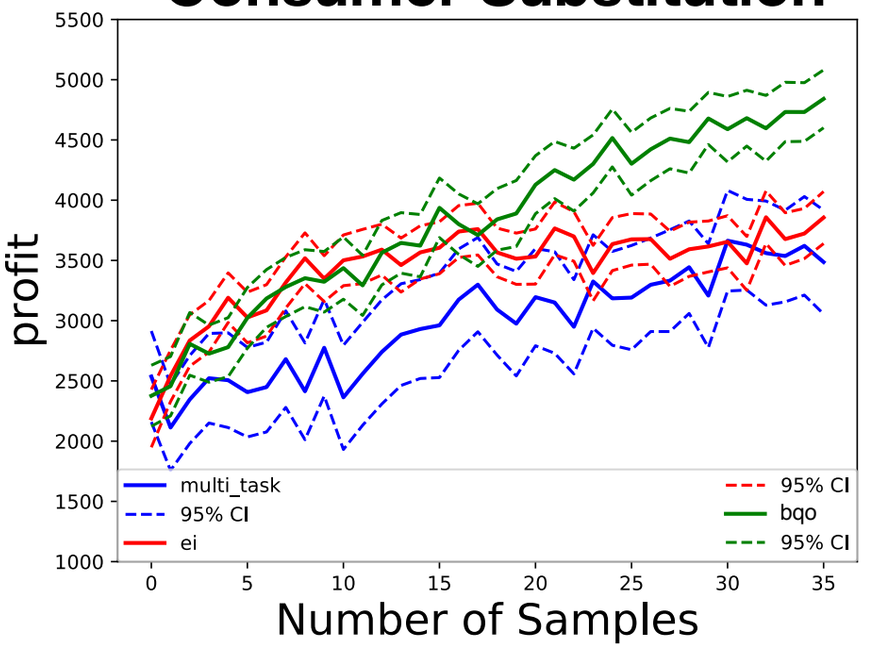
New York City's Citi Bike System



Hyperparameter Tuning in Convolutional Neural Networks



Newsvendor Problem under Dynamic Consumer Substitution



Conclusion

- Bayesian Quadrature Optimization works very well for objectives that are sums or integrals of expensive-to-evaluate integrands.

- It is derived from a conceptual one-step optimality analysis.

- The method is consistent when the objective is a finite sum.

Thank you! Any Questions?

Toscano-Palmerin & Frazier, submitted to Operations Research
Bayesian Optimization with Expensive Integrands