Uncertainty Quantification
Lecture 3: UQ for Emulators

http://arxiv.org/abs/1303.3079

Philip B. Stark
(joint work with Jeffrey C. Regier)

Department of Statistics
University of California, Berkeley

L'école Doctorale–Winter School 2017
Les Diablerets, Switzerland
5–8 February 2017

Emulators, Surrogate functions, Metamodels

Common to approximate "expensive" functions from few values.

Expense computational or real (e.g., experiment).

- Kriging
- Multivariate Adaptive Regression Splines (MARS)
- Projection Pursuit Regression
- Polynomial Chaos Expansions
- Gaussian process models (GP)
- Neural networks
- etc.

Noiseless non-parametric function estimation

- True f infinite-dimensional, on possibly high-dimensional domain.
- Observe only n samples from f.
- Estimating f is grossly underdetermined problem.
- Usual context is scientific problem involving values of f where it was not observed.

Common context

Part of larger problem in uncertainty quantification (UQ):

- Real-world phenomenon
- Physics description of phenomenon
- Theoretical simplification/approximation of the physics
- Numerical solution of the approximation f
- Emulation of the numerical solution of the approximation $\hat{f}$
- Calibration to noisy data
- "Inference"

HEB Models

High dimensional domain, Expensive, Black-box.

- Climate models (Covey et al., 2011: 21–28-dimensional domain 1154 simulations, Kriging and MARS)
- Car crashes (Aspenberg et al., 2012: 15-dimensional domain; 55 simulations; polynomial response surfaces and neural networks).
- Chemical reactions (Holena et al., 2011: 20–30-dimensional domain, boosted surrogate models; Shorter et al., 1999: 46-dimensional domain)
- Aircraft design (Srivastava et al., 2004: 25-dimensional domain, 500 simulations, response surfaces and Kriging; Koch et al., 1999: 22-dimensional domain, minutes per run, response surfaces and Kriging; Booker et al., 1999: 31-dimensional domain, minutes to days per run, Kriging).
- Electric circuits (Bates et al., 1996: 60-dimensional domain; 216 simulations; Kriging).

How accurate are emulators?

- High-consequence decisions are made on the basis of emulators.
- How accurate are they in practice?
- How can the accuracy be estimated reliably, measured or bounded?
- How many training data are needed to ensure that an emulator is accurate?

Common strategies

- For Bayesian emulators, common to use the posterior distribution to measure uncertainty (Tebaldi & Smith, 2005)
- Also common to measure error using observations not used to train the emulator (Fang et al., 2006)
- Required conditions generally cannot be verified or known to be false.
- Posterior depends on prior and likelihood, but inputs are generally fixed parameters, not random.
- Validation on hold-out observations relevant if the error at the held-out observations is representative of the error everywhere. Observations not usually IID; values of f not IID.

Constraints are key

- Without constraints on f, no reliable way to extrapolate to values of f at unobserved inputs: completely indeterminate.
- Need f to have some kind of regularity; does not typically come from the problem.
- Uncertainty estimates are driven by assumptions about f.
- Stronger assumptions $\to$ smaller uncertainties.
- What do the data justify?
- How can we avoid foolhardy optimism?

Lipschitz bound

Use absolute condition number aka Lipschitz constant:

Given a metric d on $\mathsf{dom}(g)$, best Lipschitz constant K for g is

$$K(g) \equiv \sup \left\{ \frac{g(v) - g(w)}{d(v, w)} : v, w \in \mathsf{dom}(g) \text{ and } v \neq w \right\}. \quad (1)$$

If $f \notin \mathcal{C}[0, 1]^p$, then $K(f) \equiv \infty$.

What's the problem?

- If we knew f, we could emulate it perfectly—by f.
- Require emulator $\hat{f}$ to be computable from the observations, without relying on any other information about f.
- If we knew that the Lipschitz constant of f is K, could guarantee of some level of accuracy.
- All else equal, the larger K is, the more difficult it is to guarantee that an approximation of f is accurate.

What do we know about K?

Observations $f|_X$ impose a lower bound on K (but no upper bound).

$\exists\ \hat{f}$, computable from the data $f|_X$, guaranteed to be accurate throughout the domain of f—no matter what f is—provided f agrees with the observations $f|_X$ and has a Lipschitz constant not greater than the observed lower bound on K?

Minimax formulation: Information-Based Complexity

- potential error: minimax error of emulators over the set $\mathcal{F}$ of functions that agree with data & have Lipschitz constant no greater than the lower bound, as function over $\mathsf{dom}(f)$
- maximum potential error: supremum of potential error over $\mathsf{dom}(f)$
- For known K, finding potential error is standard problem in information-based complexity.
- K is unknown since f is only partially observed. We bound potential error using a lower bound for K computed from data.

Sketch of results

- Lower bound on number of additional observations possibly necessary to "learn" f w/i ϵ.
- Application to Community Atmosphere Model: n required could be astronomical.
- Two lower bounds on the maximum potential error for approximating f from a fixed set of observations: empirical, and as a fraction of the unknown K.
- Conditions under which a constant emulator has smaller maximum potential error than best emulator trained on the actual observations. Conditions hold for the CAM simulations.
- Use sampling to estimate quantiles and mean of the potential error across the domain. For CAM, moderate quantiles are a large fraction of maximum.

Notation and problem formulation

f: fixed unknown real-valued function on $[0,1]^p$
$\mathcal{C}[0,1]^p$: real-valued continuous functions on $[0,1]^p$
$\mathsf{dom}(g)$: domain of function g
$g|_D$: restriction of g to $D \subset \mathsf{dom}(g)$
$f|_X$: data, observations of f on X
$\hat{f}$: emulator based on $f|_X$, but no other information about f
$\|h\|_\infty \equiv \sup_{w \in \mathsf{dom}(h)} |h(w)|$
d: a metric on $\mathsf{dom}(g)$
$K(g)$: best Lipschitz constant for g

$$\mathcal{F}_\kappa(g) \equiv \{h \in \mathcal{C}[0,1]^p : K(h) \leq \kappa \text{ and } h|_{\mathsf{dom}(g)} = g\}.$$

$\mathcal{F}_\infty(f|_X)$ is the space of functions in $\mathcal{C}[0,1]^p$ that fit the data.

potential error of $\hat{f} \in \mathcal{C}[0,1]^p$ over the set of functions $\mathcal{F}$:

$$\mathcal{E}(w; \hat{f}, \mathcal{F}) \equiv \sup \left\{ |\hat{f}(w) - g(w)| : g \in \mathcal{F} \right\}.$$

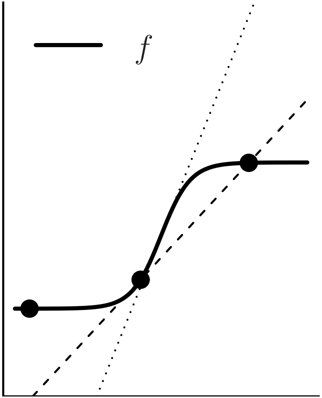
maximum potential error of $\hat{f} \in \mathcal{C}[0,1]^p$ over the set of functions $\mathcal{F}$:

$$\mathcal{E}(\hat{f}, \mathcal{F}) \equiv \sup_{w \in [0,1]^p} \mathcal{E}(w; \hat{f}, \mathcal{F}) = \left\{ \|\hat{f} - g\|_\infty : g \in \mathcal{F} \right\}.$$

Maximum potential error

- Example of worst-case error in IBC.
- The uncertainty $\hat{f}$ is $\mathcal{E}(\hat{f}, \mathcal{F}_\infty(f|_X)$.
- Presumes $f \in \mathcal{C}[0,1]^p$.
- If $f \notin \mathcal{C}[0,1]^p$, $\hat{f}$ could differ from f by more.
- We lower-bound uncertainty of the best possible emulator of f, under optimistic assumptions about the regularity of f.
- maximum potential error is infinite unless f has more regularity than continuity.

Let $K \equiv K(f)$ and $\hat{K} \equiv K(f|_X)$. Because $X \subset [0,1]^p$, $\hat{K} \leq K$.



Dotted line is tangent to f where f attains its Lipschitz constant: slope K. The dashed line is the steepest line that intersects any pair of observations: slope $\hat{K} \leq K$.

More notation

$$\mathcal{F}_\kappa \equiv \mathcal{F}_\kappa(f|_X)$$

and

$$\mathcal{E}_\kappa(\hat{f}) \equiv \mathcal{E}(\hat{f}, \mathcal{F}_\kappa).$$

radius of $\mathcal{F} \subset \mathcal{C}[0,1]^p$ is

$$r(\mathcal{F}) \equiv \frac{1}{2} \sup \{\|g-h\|_\infty : g, h \in \mathcal{F}\}.$$

$$\mathcal{E}_\kappa(\hat{f}) \geq r(\mathcal{F}_\kappa). \tag{2}$$

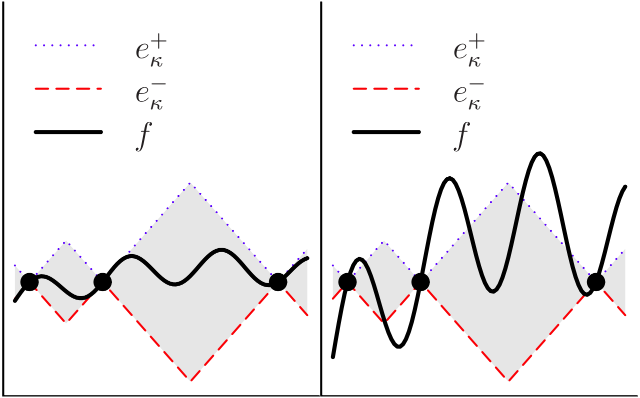
Equality holds for the emulator that "splits the difference":

$$f_\kappa^\star(w) \equiv \frac{1}{2}\left[\inf_{g \in \mathcal{F}_\kappa} g(w) + \sup_{g \in \mathcal{F}_\kappa} g(w)\right]$$

That is, for all emulators $\hat{f}$ that agree with f on X,

$$\mathcal{E}_\kappa(\hat{f}) \geq \mathcal{E}_\kappa(\hat{f}_\kappa^*) \equiv \mathcal{E}_\kappa^* :$$

$f_\kappa^\star$ is a minimax (over $f \in \mathcal{F}_\kappa$) for infinity-norm error.



$\hat{K} = 0$; optimal interpolant $f_\kappa^\star$ is constant. Left panel: $\kappa = K$. Right panel: $\kappa < K$. If $\kappa \geq K$ then $e_\kappa^- \leq f \leq e_\kappa^+$, and, equivalently, $f \in \mathcal{F}_\kappa$.

Define

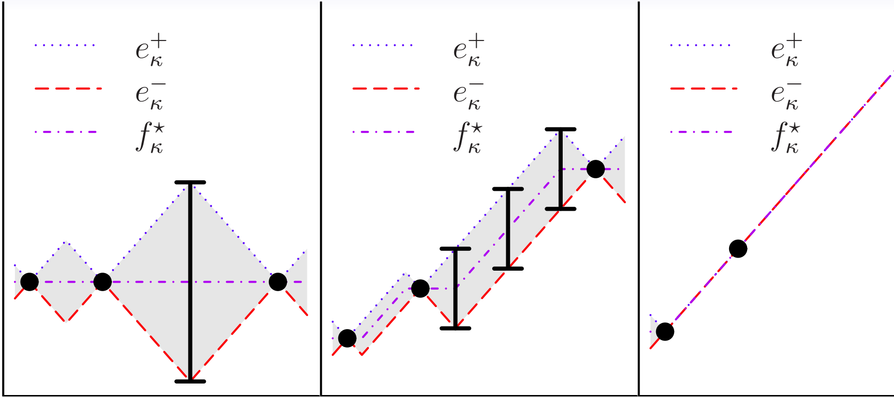
$$e_\kappa^+(w) \equiv e_{f,X,\kappa}^+(w) \equiv \min_{x \in X} \left[f(x) + \kappa d(x,w) \right],$$

$$e_\kappa^-(w) \equiv e_{f,X,\kappa}^-(w) \equiv \max_{x \in X} \left[f(x) - \kappa d(x,w) \right],$$

and

$$e_\kappa^\star(w) \equiv e_{f,X,\kappa}^\star(w) \equiv \frac{1}{2} \left[e_{f,X,\kappa}^+(w) - e_{f,X,\kappa}^-(w) \right].$$

$e_\kappa^\star(w)$ is minimax error at w: smallest (across emulators $\hat{f}$) maximum (across functions g) error at the point $w \in [0,1]^p$ is $e_\kappa^\star(w)$.



Black error bars are twice the maximum potential error over $\mathcal{F}_\kappa$. The succession of panels shows that as the slope between observations approaches κ, $\mathsf{e}^\star(w)$ approaches 0 for points w between observations, and the maximum potential error over $\mathcal{F}_\kappa$ decreases.

Bounds on the number of observations

Fix “tolerable error” $\epsilon > 0$

If $\left\| \hat{f}|_A - g|_A \right\|_\infty \leq \epsilon$, then $\hat{f}$ ϵ-approximates g on A. If $A = \mathsf{dom}(g)$, then $\hat{f}$ ϵ-approximates g.

If $\mathcal{F}$ is a non-empty class of functions with common domain D, then $\hat{f}$ ϵ-approximates $\mathcal{F}$ on $A \subset D$ if $\forall g \in \mathcal{F}$, $\hat{f}$ ϵ-approximates g on A. If $A = D$, then $\hat{f}$ ϵ-approximates $\mathcal{F}$.

ϵ-approximates

$\hat{f}$ ϵ-approximates $\mathcal{F}$ if and only if the maximum potential error of $\hat{f}$ on $\mathcal{F}$ does not exceed ϵ.

Since $\hat{K}$ is the observed variation of f on X, a useful value of ϵ would typically be much smaller than $\hat{K}$. (Otherwise, we might just as well take $\hat{f}$ to be a constant.)

For fixed $\epsilon > 0$, and $Y \subset \mathsf{dom}(f)$, Y is ϵ-adequate for f on A if $f_K^\star$ ϵ-approximates $\mathcal{F}_K(f|_Y)$ on A. If $A = \mathsf{dom}(f)$, then Y is ϵ-adequate for f.

$B(x, \delta)$: open ball in $\mathbb{R}^p$ centered at x with radius δ.

$$N_f \equiv \min\{\#Y : Y \text{ is } \epsilon\text{-adequate for } f\},$$

where $\#Y$ is the cardinality of Y.

The minimum potential computational burden is

$$M \equiv \max\{N_g : g \in \mathcal{F}_K\}.$$

Over all experimental designs Y, M is the smallest number of data to guarantee that maximum error of the best emulator based on those data is not larger than ϵ.

Upper bound on N_f

For each $x \in X$, $f_K^\star$ ϵ-approximates $\mathcal{F}_K(f|_K)$ on (at least) $B(x, \epsilon/K)$. Thus, $f_K^\star$ ϵ-approximates $\mathcal{F}_K$ on $\bigcup_{x \in X} B(x, \epsilon/K)$. Hence, the cardinality of any $Y \subset [0,1]^p$ for which

$$V \equiv \left\{ B\left(x, \frac{\epsilon}{K}\right) : x \in Y \right\} \supset [0,1]^p$$

is an upper bound on N_f.

In ℓ_∞, $[0,1]^p$ can be covered by $\left\lceil \frac{K^+}{2\epsilon} \right\rceil^p$ balls of radius ϵ/K^+.

Lower bound on N_f

- Can happen that $f^\star_{\hat{K}}$ ϵ-approximates $\mathcal{F}_K$ on regions of the domain not contained in $\cup_{x \in X} B(x, \epsilon/K)$.
- If f varies on X, then for a function g to agree with f at the observations requires g to vary too.
- Fitting the data"spends" some of g's Lipschitz constant: can't get as far away from f as it could if f_X were constant.
- Can quantify to find lower bounds for M.

Define $\bar{\gamma} \equiv \arg\min_{\gamma \in \mathbb{R}} \sum_{x \in X} |f(x) - \gamma|^p$.

Let $X^+ \equiv \{x \in X : f(x) \geq \bar{\gamma}\}$ and let $X^- \equiv \{x \in X : f(x) < \bar{\gamma}\}$.
Let

$$Q_+ \equiv \bigcup_{x \in X^+} \left\{ B\left(x, \frac{f(x) - \bar{\gamma}}{\hat{K}}\right) \bigcap [0,1]^p \right\}$$

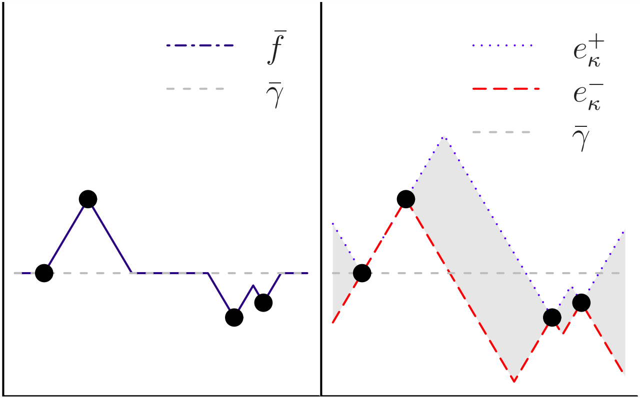
and

$$Q_- \equiv \bigcup_{x \in X^-} \left\{ B\left(x, \frac{\bar{\gamma} - f(x)}{\hat{K}}\right) \bigcap [0,1]^p \right\}.$$

Then $Q_+ \cap Q_- = \emptyset$

Define

$$\bar{f} : [0,1]^p \to \mathbb{R}$$
$$w \mapsto \begin{cases} e^-_{\hat{K}}(w), & w \in Q_+ \\ e^+_{\hat{K}}(w), & w \in Q_- \\ \bar{\gamma}, & \text{otherwise.} \end{cases}$$



$\bar{f}$ (left panel) is comprised of segments of $e_{\hat{K}}^+$, $e_{\hat{K}}^-$ and the constant $\bar{\gamma}$ (right panel). $\bar{f}$ constant over roughly half of the domain. No function between $e_{\hat{K}}^-$ and $e_{\hat{K}}^+$ (inclusive) is constant over a larger fraction of the domain.

Result 1

μ: Lebesgue measure. $\bar{Q} \equiv [0,1]^p \setminus (Q_+ \cup Q_-)$.

$$\mu(\bar{Q}) \geq 1 - \sum_{x \in X} \mu\left(B\left(x, |f(x) - \bar{\gamma}|/\hat{K}\right)\right).$$

$C_2 \equiv \frac{\pi^{p/2}}{\Gamma(p/2+1)}$ and $C_\infty \equiv 2^p$. For $q \in \{2, \infty\}$,

$$\mu(\bar{Q}) \geq 1 - C_q \sum_{x \in X} \left(|f(x) - \bar{\gamma}|/\hat{K}\right)^p.$$

If $\exists x \in X$ for which $\{x\}$ is ϵ-adequate for f on $A \subset \bar{Q}$, then $\mu(A) \leq \mu(B(0, \epsilon/\hat{K}))$.

$$M \geq \left\lceil \frac{\mu(\bar{Q})}{\mu(B(0, \epsilon/\hat{K}))} \right\rceil \geq \left\lceil \epsilon^{-p} \left[\frac{\hat{K}^p}{C_q} - \sum_{x \in X} |f(x) - \bar{\gamma}|^p \right] \right\rceil. \quad (3)$$

PCMDI

- Program for Climate Model Diagnosis & Intercomparison (PCMDI) at LLNL: 1154 climate simulations using the Community Atmosphere Model (CAM).
- $p = 21$ parameters scaled so that $[0, 1]$ has all plausible values.
- f is global average upwelling longwave flux (FLUT) approximately 50 years in the future.
- Each run took several days on a supercomputer.
- PCDMI used several approaches to choose $X \subset [0, 1]^p$: Latin hypercube, one-at-a-time, and random-walk multiple-one-at-a-time.
- 1154 simulations total.

$\bar{\gamma} = 232.77$; $\hat{K} = 14.20$ for $q = 2$:

$$M \geq \left\lceil \epsilon^{-21} \left[\frac{1.57 \times 10^{24}}{0.0038} - 6.81 \times 10^{24} \right] \right\rceil > \epsilon^{-21} \times 10^{26}.$$

If ϵ is 1% of $\hat{K}$, then $M \geq 10^{43}$.

Even if ϵ is 50% of $\hat{K}$, $M > 10^{8}$. For $q = \infty$, $\hat{K} = 34.68$; in that case

$$M \geq \left\lceil \epsilon^{-21} \left[\frac{2.19 \times 10^{32}}{2^{21}} - 6.81 \times 10^{25} \right] \right\rceil > \epsilon^{-21} \times 10^{25}.$$

Lower bounds on maximum potential error

- Two lower bounds on the maximum potential error $\mathcal{E}_K^*$ for fixed X: absolute, and as a fraction of unknown K.
- Bound as fraction of K shows that when a statistic—calculable from the observations—exceeds a calculable threshold, the maximum potential error is no less than the maximum potential error from one observation at the centroid.
- Observing f for all $x \in X$ was wasteful: one observation would have been better.
- For LLNL CAM runs, both bounds are large.

Result 2

Theorem

$$\mathcal{E}_K(\hat{f}) \geq \sup e^\star_{\hat{K}}.$$

$\sup e^\star_{\hat{K}}$, a statistic calculable from data $f|_X$, is a lower bound on the maximum potential error for any emulator $\hat{f}$ based on the observations $f|_X$.

Result 3: Scaling Lemma

Lemma

For any λ, if $\sup e^{\star}_{\hat{K}} \geq \lambda \hat{K}$, then $\mathcal{E}_K(\hat{f}) \geq \lambda K$.

Maximum potential error from 1 observation

Work in ℓ_∞: $d(v, w) = \|v - w\|_\infty$.

$z \equiv (1/2, \ldots, 1/2)$, the centroid of $[0, 1]^p$.

$\hat{g} \in \mathcal{F}_\infty(f|_{\{z\}})$ is constant function $\hat{g}(w) \equiv f(z)$, $\forall w \in [0, 1]^p$.

ℓ_∞ distance from z to any boundary point of $[0, 1]^p$ is $1/2$, so

$$\mathcal{E}_K(\hat{g}, \mathcal{F}_K(f|_{\{z\}})) = \frac{K}{2}.$$

Result 4

Let $W \subset [0,1]^p$ be finite and $c \in \mathbb{R}$. Suppose $f|_W = c$. Let $\hat{h} \in \mathcal{F}_\infty(f|_W)$. By examining the corners of the domain, it follows that if $|W| < 2^p$,

$$\mathcal{E}_K(\hat{h}, \mathcal{F}_K(f|_W)) \geq \frac{K}{2}.$$

If f is constant on W, any emulator based on fewer than 2^p observations of f will have at least $K/2$ maximum potential error.

Making 2^p observations of f is intractable for CAM and many other applications.

Result 5

Theorem

If $\sup e^{\star}_{\hat{K}} \geq \hat{K}/2$, then

$$\mathcal{E}_K(\hat{f}) = \mathcal{E}_K(\hat{f}, \mathcal{F}_K(f|_X)) \geq \frac{K}{2} \geq \mathcal{E}_K(\hat{g}, \mathcal{F}_K(f|_{\{z\}})).$$

If $\sup e^{\star}_{\hat{K}} \geq \hat{K}/2$, no $\hat{f}$ based on $f|_X$ has smaller maximum potential error than the constant emulator based on one observation.

CAM: Upper bound from non-adjacent corners in ℓ_∞.

Theorem

$$\sup e^\star_{\hat{K}} \leq \frac{1}{2}\left\{\min_{x\in X}\left[f(x)+\hat{K}\tilde{d}(x)\right]-\max_{x\in X}\left[f(x)-\hat{K}\tilde{d}(x)\right]\right\}.$$

$\sup e^\star_{\hat{K}} \leq 20.95$ for the CAM dataset.

CAM: Lower bounds from corners in ℓ_∞.

Clearly

$$\sup e^\star_{\hat{K}} \geq \max\left\{e^\star_{\hat{K}}(w) : \forall w \in \{0,1\}^p\right\}.$$

Essentially sharp for the CAM dataset.

Divide $[0,1]^p$ into 2^p hypercubes $\{R_i\}_{i=1}^{2^p}$ with edge-length $1/2$, disjoint interiors, each containing a different corner of $[0,1]^p$

Because X contains only 1154 points, most R_i do not contain any $x \in X$.

The bounds are tight for CAM

For the CAM dataset, one corner r_j attains $e^\star_{\hat{K}}(r_j) = 20.95$.

So, $e^\star_{\hat{K}}$ attains the upper bound established in the previous section, and $\sup e^\star_{\hat{K}} = 20.95$.

Implications for CAM

Because $\sup e^{\star}_{\hat{K}} = 20.95 \geq 17.34 = \hat{K}/2$, $\mathcal{E}_K(\hat{f}) \geq K/2$ for any interpolation $\hat{f}$.

Maximum potential error would have been no greater had we just observed f once, at z, and predicted $\hat{f}(w) = f(z)$ for all $w \in [0,1]^p$.

Extensions

- Looked at maximum uncertainty over all $w \in [0,1]^p$.
- Important in some applications; in others, maybe less interesting than the fraction of $[0,1]^p$ where uncertainty is large.
- Can estimate the fraction of $[0,1]^p$ for which $e^* \geq \epsilon > 0$ by sampling.
- Draw $w \in [0,1]^p$ at random and evaluate e^* at each selected point.
- Yields binomial lower confidence bounds for the fraction of $[0,1]^p$ where uncertainty is large, and confidence bounds for quantiles of the potential error.

CAM: bounds on percentiles of error

	95% lower confidence bound			
norm	lower quartile	median	upper quartile	average
Euclidean	1.454	1.596	1.731	1.595
supremum	0.649	0.717	0.782	0.715

Error of minimax emulator $f^\star_{\hat{K}}$ of CAM model from 1154 LLNL observations. Column 1: metric d used to define the Lipschitz constant. Columns 2–4: Binomial lower confidence bounds for quartiles of the pointwise error. Column 5: 95% lower confidence bound for the integral of the pointwise error over the entire domain $[0, 1]^p$. Columns 2–5 are expressed as multiple of $\hat{K}/2$. Based on 10,000 random samples.

Conclusions

- In some problems, every emulator based on any tractable number of observations of f has large maximum potential error (and the potential error is large over much of the domain), even if f is no less regular than it is observed to be.
- Can find sufficient conditions under which all emulators are potentially substantially incorrect.
- Conditions depend only on the observed values of f; can be computed from the same observations used to train an emulator, at small incremental cost.
- Conditions are sufficient but not necessary: f could be less regular than any finite set of observations reveals it to be.
- It is not possible to give necessary conditions that depend only on the data.
- Conditions seem to hold for problems with large societal interest.

- Reducing the potential error of emulators in HEB problems requires either more information about f (knowledge, not merely assumptions), or changing the measure of uncertainty—changing the scientific question.
- Both tactics are application-specific: the underlying science dictates the conditions that actually hold for f and the senses in which it is useful to approximate f.
- Not clear that emulators help address the most important questions.
- Approximating f pointwise rarely ultimate goal; most properties of f are nuisance parameters.
- Important questions about f might be answered more directly.
- Some research questions cannot be answered through simulation at present.
- Employing complex emulators and massive computational is a distraction.