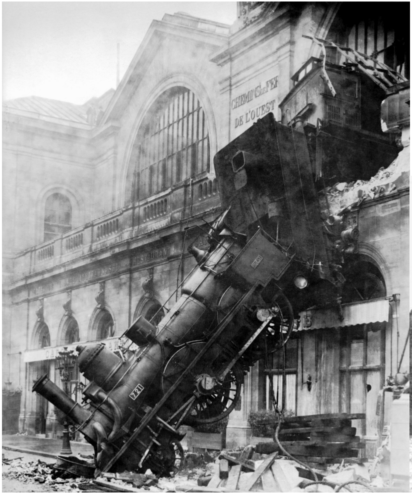
Uncertainty Quantification for Emulators
http://arxiv.org/abs/1303.3079

Philip B. Stark and Jeffrey C. Regier

Department of Statistics
University of California, Berkeley

Dipartimento di Fisica e Astronomia
Università di Bologna
Bologna, Italy
5 June 2013

Why Uncertainty Quantification Matters



??

James Bashford / AP

Why Uncertainty Quantification Matters

NASA

Reuters / Japan TSB

Emulators, Surrogate functions, Metamodels

Try to approximate a function f from few samples when evaluating f expensive: computational cost or experiment.

Emulators are essentially interpolators/smoothers

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

Noiseless non-parametric function estimation

Estimate f on domain $\mathrm{dom}(f)$ from $\{f(x_1), \ldots, f(x_n)\}$

- f infinite-dimensional. $\mathrm{dom}(f)$ typically high-dimensional.
- Observe only $f|_X$, where $X = \{x_1, \ldots, x_n\}$. No noise.
- Estimating f is grossly underdetermined problem (worse with noise).
- Usual context: question that requires knowing $f(x)$ for $x \notin X$

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: *H*igh dimensional domain, *E*xpensive, *B*lack-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, NN)
- 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)

Emulator Accuracy Matters

- 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 to estimate accuracy

Bayesian Emulators (GP, Kriging, . . .)

- Use the posterior distribution (Tebaldi & Smith 2005)
- Posterior depends on prior and likelihood, but inputs are generally fixed parameters, not random.

Others

- Using holdout data (Fang et al. 2006)
- Relevant only if the error at the held-out data is representative of the error everywhere. Data not usually IID; values of f not IID.

Required conditions generally unverifiable or known to be false.

So, what to do?

- Standard methods can be misleading when the assumptions don't hold— and usually no reason for the assumptions to hold.
- Is there a more rigorous way to evaluate the accuracy?
- Is there a way that relies only on the observed data?

Constraints are mandatory

- Uncertainty estimates are driven by *assumptions* about f.
- Without constraints on f, no reliable way to extrapolate to values of f at unobserved inputs: completely uncertain.
- Stronger assumptions $\rightarrow$ smaller uncertainties.
- What's the most optimistic assumption the data justify?

(Best) Lipschitz constant

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

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

If $f \notin \mathcal{C}(\mathrm{dom}(f))$, 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 data, without relying on any other information about f.
- If we knew $K(f)$, could guarantee *some* level of accuracy for $\hat{f}$.
- All else equal, the larger $K(f)$ is, the harder to guarantee that $\hat{f}$ is accurate.

How bad *must* the uncertainty be?

- Data $f|_X$ impose a lower bound on $K(f)$ (but no upper bound): Data *require* some lack of regularity.
- Is there any $\hat{f}$ guaranteed to be close to f—no matter what f is—provided f agrees with $f|_X$ and is not less regular than the data require?

Minimax formulation: Information-Based Complexity (IBC)

- *potential error at w*: minimax error of emulators $\hat{f}$ over the set $\mathcal{F}$ of functions g that agree with data & have $K(g)$ constant no greater than the lower bound, at $w \in \mathrm{dom}(f)$.
- *maximum potential error*: sup of potential error over $w \in \mathrm{dom}(f)$.
- For known K, finding potential error is standard IBC problem.
- But $K(f)$ is unknown: Bound potential error using a lower bound for $K(f)$ computed from data.

Sketch of results

- Lower bound on additional observations possibly necessary to estimate f w/i ϵ.
- Application to Community Atmosphere Model (CAM): required n could be ginormous.
- Lower bounds on the max 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.
- Sampling to estimate quantiles and mean of the potential error over dom(f). For CAM, moderate quantiles are a large fraction of maximum.

Notation

f: fixed unknown real-valued function on $[0,1]^p$

$\mathcal{C}[0,1]^p$: real-valued continuous functions on $[0,1]^p$

$\text{dom}(g)$: domain of the function g

$g|_D$: restriction of g to $D \subset \text{dom}(g)$

$f|_X$: f at the n points in X, the data

$\hat{f}$: emulator based on $f|_X$, but no other information about f

$\|h\|_\infty \equiv \sup_{w \in \text{dom}(h)} |h(w)|$

d: a metric on $\text{dom}(g)$

$K(g)$: best Lipschitz constant for f (using metric d)

More notation

- κ-smooth interpolant of g:

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

  $\mathcal{F}_\infty(f|_X)$ 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* $\mathcal{F}$:

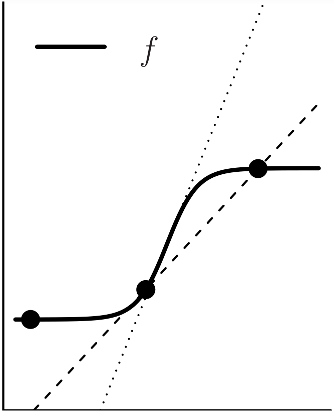
$$\mathcal{E}(w; \hat{f}, \mathcal{F}) \equiv \sup \left\{ |\hat{f}(w) - g(w)| : g \in \mathcal{F} \right\}.$$

- *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.
- "Real" uncertainty of $\hat{f}$ is $\mathcal{E}(\hat{f}, \mathcal{F}_\infty(f|_X))$.
- Presumes $f \in \mathcal{C}[0,1]^p$.
- Maximum potential error is infinite unless f has more regularity than continuity.
- 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 assumption that $K = K(f) = \hat{K} \equiv K(f|_X) \leq K(f)$



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

More notation

- $\mathcal{F}_\kappa \equiv \mathcal{F}_\kappa(f|_X)$
- $\mathcal{E}_\kappa(\hat{f}) \equiv \mathcal{E}(\hat{f}, \mathcal{F}_\kappa)$
- *radius* of $\mathcal{F} \subset \mathcal{C}[0,1]^p$ is

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

First result

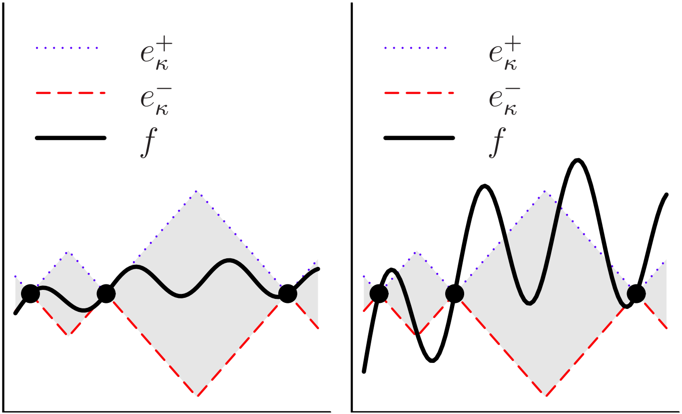
$$\mathcal{E}_\kappa(\hat{f}) \geq r(\mathcal{F}_\kappa). \qquad (2)$$

Equality holds for the emulator that “splits the difference”:

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

For all emulators $\hat{f}$ that agree with f on X,

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



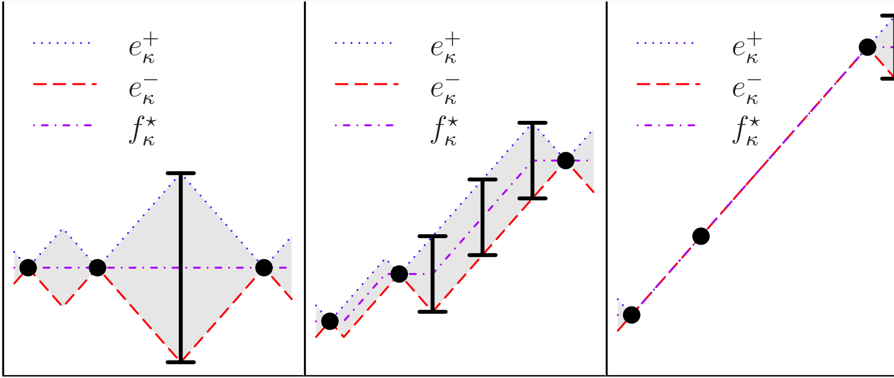
Left panel: $\kappa = K$. Right panel: $\kappa < K$.
If $\kappa \geq K$ then $e_\kappa^- \leq f \leq e_\kappa^+$, so $f \in \mathcal{F}_\kappa$.

Constructing e^-, e^+, and $e^\star$

Define

- $e^+_\kappa(w) \equiv \min_{x \in X} \left[f(x) + \kappa d(x, w)\right]$
- $e^-_\kappa(w) \equiv \max_{x \in X} \left[f(x) - \kappa d(x, w)\right]$
- $e^\star_\kappa(w) \equiv \frac{1}{2}\left[e^+_{f,X,\kappa}(w) - e^-_{f,X,\kappa}(w)\right]$

$e^\star_\kappa(w)$ is minimax error at w:
smallest (across emulators $\hat{f}$) maximum (across functions g) error at the point w



Black error bars are twice the maximum potential error over $\mathcal{F}_\kappa$. As the slope between observations approaches κ, $e^\star(w)$ approaches 0 for points w between observations, and the maximum potential error over $\mathcal{F}_\kappa$ decreases.

Lower bounds on n

- 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 $A = \mathrm{dom}(g)$, 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 and tolerable error

$\hat{f}$ 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.)

Minimum potential computational burden

- For fixed $\epsilon > 0$, and $Y \subset \mathrm{dom}(f)$, Y is *ϵ-adequate* for f on A if $f_K^\star$ ϵ-approximates $\mathcal{F}_K(f|_Y)$ on A. If $A = \mathrm{dom}(f)$, 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 for which the maximum error of the best emulator based on those data is guaranteed not to exceed 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.

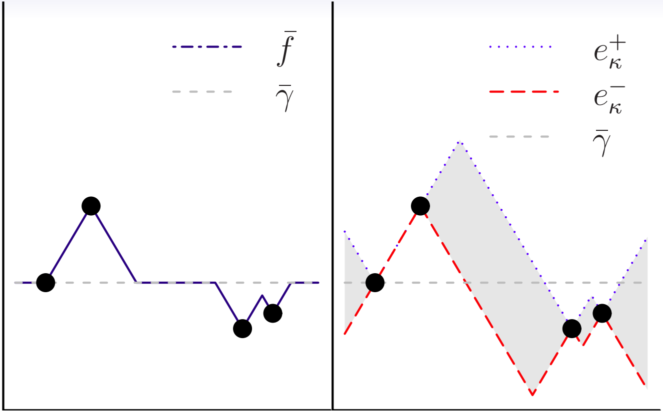
Lower bound on N_f: Heuristics

- 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.

Lower bound on N_f: Construction

Define

- $\bar{\gamma} \equiv \arg\min_{\gamma \in \mathbb{R}} \sum_{x \in X} |f(x) - \gamma|^p$.
- $X^+ \equiv \{x \in X : f(x) \geq \bar{\gamma}\}$
- $X^- \equiv \{x \in X : f(x) < \bar{\gamma}\}$.
- $Q_+ \equiv \bigcup_{x \in X^+} \left\{ B\left(x, \frac{f(x)-\bar{\gamma}}{\hat{K}}\right) \bigcap [0,1]^p \right\}$
- $Q_- \equiv \bigcup_{x \in X^-} \left\{ B\left(x, \frac{\bar{\gamma}-f(x)}{\hat{K}}\right) \bigcap [0,1]^p \right\}$
- $\bar{Q} \equiv [0,1]^p \setminus (Q_+ \cup Q_-)$.
- $\bar{f}(w) \equiv \{e^-_{\hat{K}}(w), w \in Q_+; e^+_{\hat{K}}(w), w \in Q_-; \bar{\gamma}, w \in \bar{Q}\}$.



$\bar{f}$ (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.

Potential computational burden: bounds for Lebesgue measure

- μ: Lebesgue measure.

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

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

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

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

Uncertainty Quantification Strategic Initiative–LLNL

- Uncertainty Quantification Strategic Initiative 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.
- 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.

CAM calculations

- $\bar{\gamma} = 232.77$
- For $q = 2$, $\hat{K} = 14.20$:
  $M \geq \left\lceil \epsilon^{-21} \left[\frac{1.57 \times 10^{24}}{0.0038} - 6.81 \times 10^{24} \right] \right\rceil > \epsilon^{-21} \times 10^{26}$
  If ϵ is 1% of $\hat{K}$, then $M \geq 10^{43}$.
  Even if ϵ is 50% of $\hat{K}$, $M > 10^{8}$.
- For $q = \infty$, $\hat{K} = 34.68$:
  $M \geq \left\lceil \epsilon^{-21} \left[\frac{2.19 \times 10^{32}}{2^{21}} - 6.81 \times 10^{25} \right] \right\rceil > \epsilon^{-21} \times 10^{25}$

Universal bound from the data

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 $f|_X$.

More isn't necessarily better

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 at the centroid z of $[0,1]^p$

Implications for CAM

- $\sup e^{\star}_{\hat{K}} = 20.95 \geq 17.34 = \hat{K}/2$
- Hence, $\mathcal{E}_K(\hat{f}) \geq K/2$ for every emulator $\hat{f}$.
- Maximum potential error would have been no greater had we just observed f at z and emulated by $\hat{f}(w) = f(z)$ for all $w \in [0,1]^p$.

Extensions

- Covered 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 may be a distraction.