# Bayesian Optimization with Shape Constraints

Michael Jauch
*Víctor Peña*
Department of Statistical Science
Duke University

Feb 20th, 2017

You are here

Probabilistic Numerics

Bayes Opt

Our paper

# Probabilistic Numerics

- Suppose you're using a numerical method.

- When you stop, **you don't know what the "right answer" is.**

- How can you quantify that uncertainty?

- If you buy Cox's axioms, **probability is the way to go.**

# Probabilistic Numerics

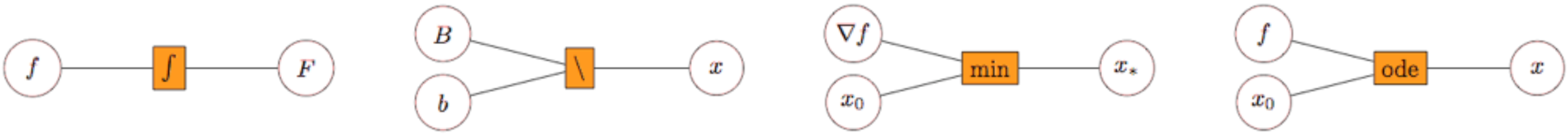
- The first application of this idea is attributed to Poincaré (1896).

- Diaconis (1988) tells us that Poincaré (1896) used Gaussian processes (GPs) for numerical interpolation (think computer models).

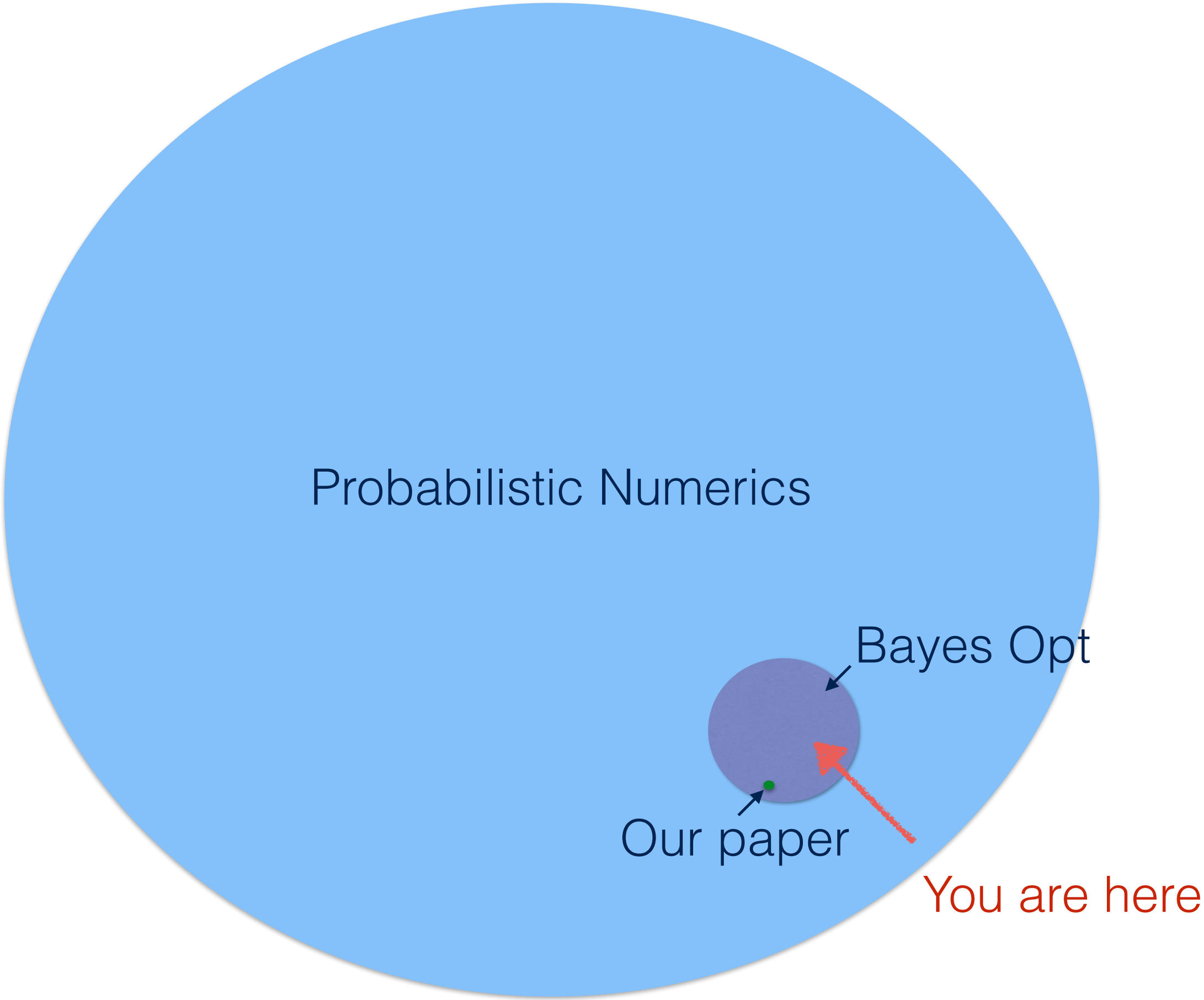
- None of the GP theory was available back then.

Henri Poincaré

# Probabilistic Numerics




- I’m new to the field and don’t know much about it.
- Hennig et al. (2015) is a great introduction.
- The community is active and growing.
- Check out probabilistic-numerics.org


Probabilistic Numerics
Bayes Opt
Our paper
You are here

# Bayesian Optimization

**What is it about?**

- **Gaussian processes** as **surrogate models for objective functions.**

**How does it work?**

1. Get started with some initial point(s).

2. Estimate where the optimum is and quantify your uncertainty.

3. If you're not "certain enough," decide where to sample next.

# Bayesian Optimization

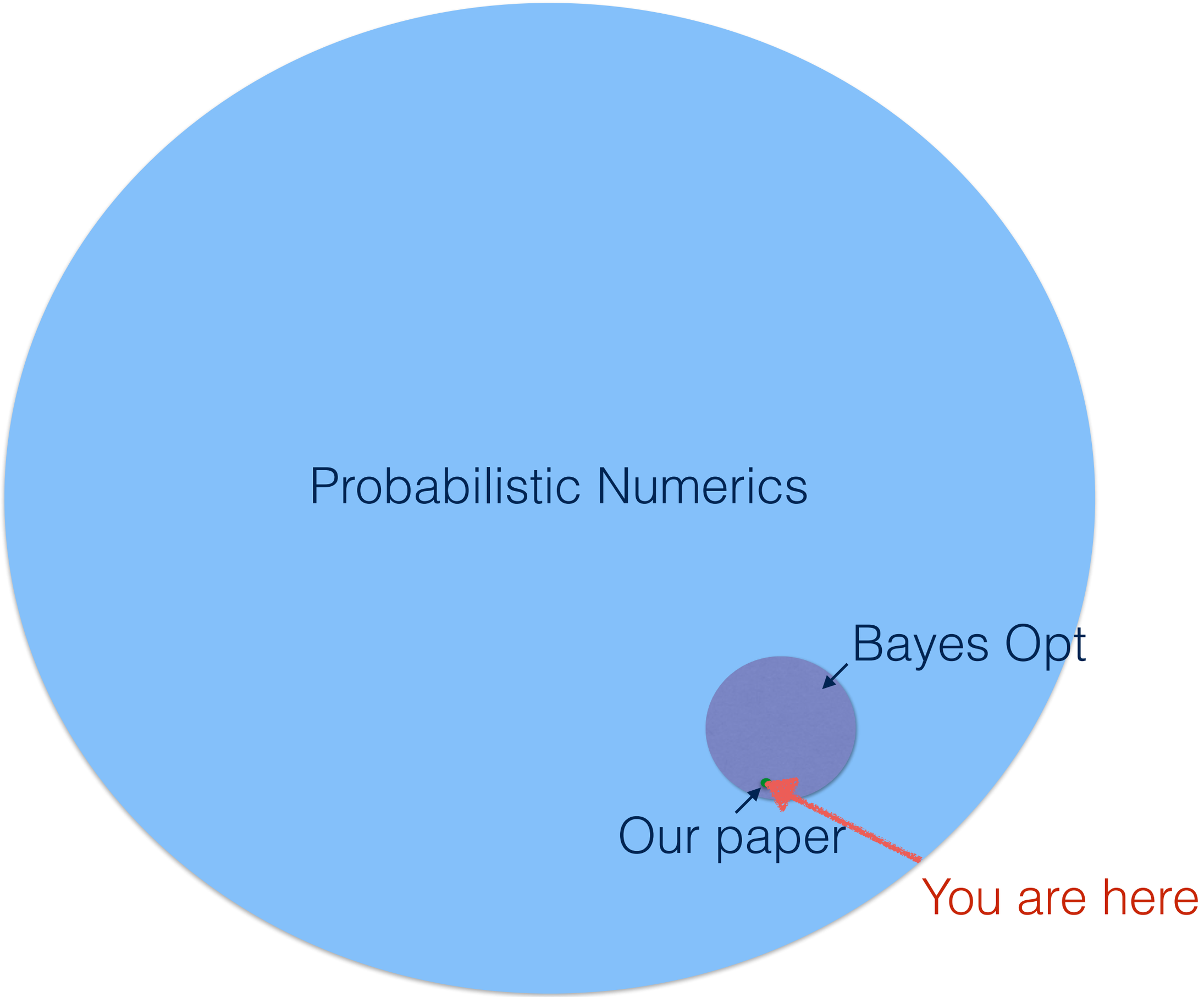
**Why would you do that?!**

- Objective functions can be **noisy**, **not available in closed-form**, and/or **expensive to evaluate**.

- GP theory tells us that we can **estimate consistently functions that are in large classes** which aren't necessarily "nice."

- Given a finite set of (possibly noisy) evaluations of the objective function, we can estimate where the optimum is and **quantify our uncertainty**. This is crucial if we can't afford to evaluate our objective function many times (one could say it's a *moral obligation*).

- If you know something about the problem, you can **use prior information**.


Probabilistic Numerics
Bayes Opt
Our paper
You are here

# Bayesian Optimization with Shape Constraints

- In some cases, **there is prior information** about the **shape** of the objective function with respect to one or more of its arguments.

- For smooth covariance functions, derivatives of GPs are **jointly Gaussian** with the observations (see e.g. Rasmussen & Williams (2006)).

- Our NIPS workshop paper:

  - Argues in favor of the appropriateness of shape constraints in a couple of explicit applications.

  - Introduces a conceptually simple (albeit approximate and not as efficient as we'd like to) way to implement them.

  - Presents a couple of toy examples with promising results.

# Derivatives of GPs

- Suppose the covariance function $k$ of the GP is smooth enough.

- Let $x, x' \in \mathbb{R}^d$ and $o, o'$ be the orders of their derivatives.

- Then,

$$\mathrm{Cov}\left(\frac{\partial^o}{\partial x_j^o} f(x), \frac{\partial^{o'}}{\partial x_k'^{o'}} f(x')\right) = \frac{\partial^{o+o'}}{\partial x_j^o x_k'^{o'}} k(x, x')$$

- And you can take a bunch of those and everything is jointly Gaussian.

# Examples of Constraints

- We can impose convexity or concavity by imposing the appropriate constraints in the derivatives of the GP.

- We also consider ***quasiconvexity***, which includes functions that are unimodal, but not necessarily convex.

- In 1D, quasiconvex functions are either (1) monotone or (2) nonincreasing up to some point, and then nondecreasing from that point onward.

    - Minor observation: This definition gives us a straightforward way to implement an accept-reject sampler in 1D, which can be extended to “slices” in higher dimensions (we do that in 2D).

# Some Challenges

- We'd like to enforce the constraints for all possible values in the input space, but we can't. We can enforce them on a grid, though.
    - **Solution:** Wang and Berger (2016) propose a (sequential) algorithm to keep on adding points until the probability that the constraint is violated is low enough.
- Covariances get big *fast*:
    - **Solution:** Thankfully, the construction of those can be vectorized (for the most part), and we can resort to approximate methods for inverting matrices (we used incomplete Choleski; there might be better ways).
- The likelihoods/predictives are high-dimensional truncated distributions.
    - **"Solution:"** In our examples, we take a "partial likelihood" approach (i.e. we ignore the constraints) for estimating the MAPs of the hyperparameters, and we use Botev (2016) to sample from the truncated posterior predictive distributions, given the (wrong) MAPs.

# Some Caveats

- Our estimates of the parameters are **off** (because they ignore the constraints), and **we underestimate our uncertainty because we're taking MAPs**. We acknowledge this is not a great solution.

- However, all our decisions, given the wrong estimates, are made via sampling from the ***correct*** truncated distributions.

  - Our acquisition function (which tells us where to sample next) is based on a MC estimate of the posterior expectation of *f* **given the constraints** (truncated normals).

  - At any iteration, our estimation of the optimum and predictive intervals are taken from the ***correct*** distribution, given the ***wrong*** estimates.

# Application: Optimal Design

- Suppose you have a (Bayesian) model and a utility function, and you want to **design an experiment** (not necessarily randomized; it can be anything).

- If you buy the Von Neumann-Morgenstern axioms, you will **choose the experiment that maximizes your utility** given your current state of knowledge.

- Unfortunately, in interesting problems one has to make decisions from **noisy samples** at a **limited number of designs.**

    - More often than not, posterior utilities aren't available in closed form and are approximated with MC methods.

    - In complicated problems, you might have a large number of possible designs.

# Optimal design & BayesOpt

- If the design variables can take on many possible values (think sample size), Bayesian optimization can help us solve the problem.

- The utility of some design variables is often convex (or at least unimodal), reflecting the **trade-off between "statistical accuracy" and sampling costs.**

# Toy Example

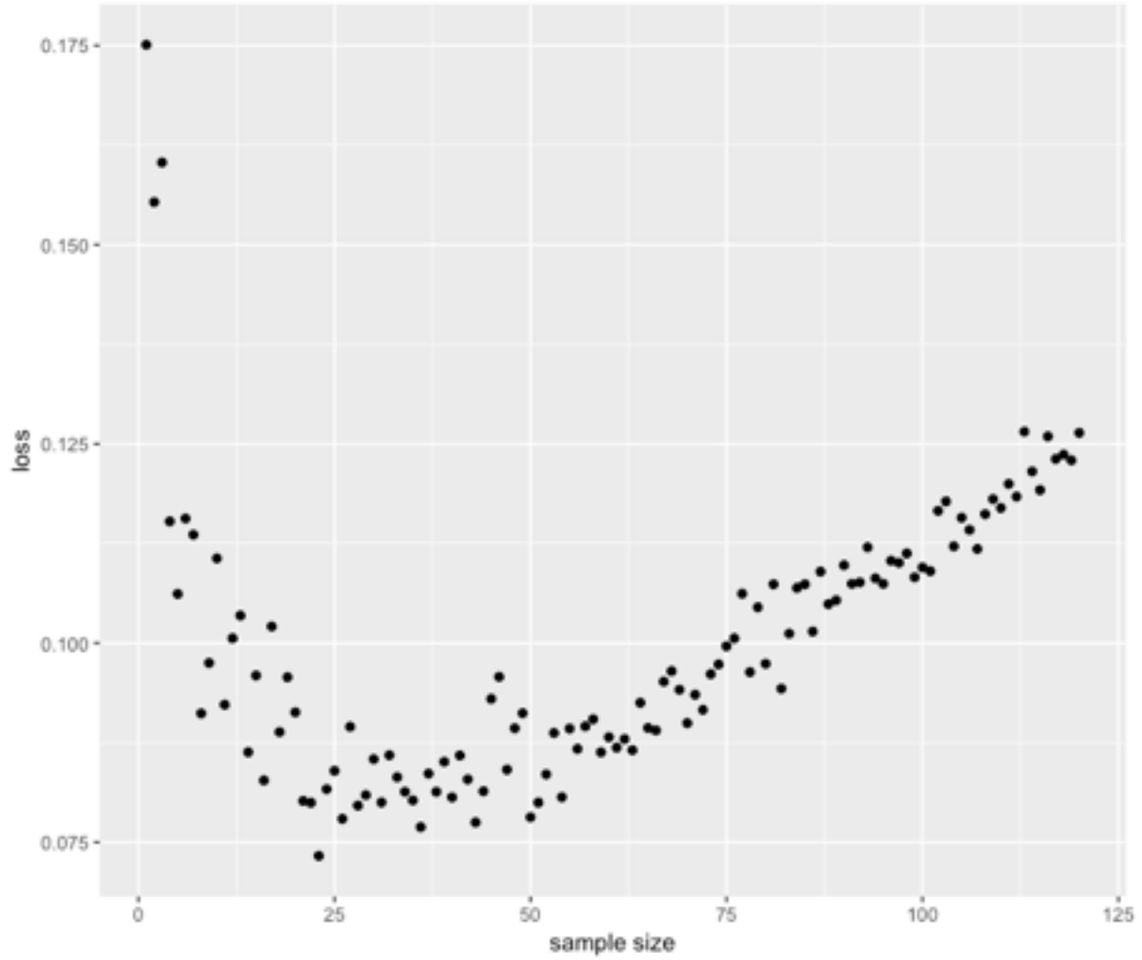
- Optimal sample size of a $\text{Binomial}(n, \theta)$ experiment where the prior on $\theta$ is an equal-weighted mixture of a Beta(3,1) and a Beta(3,3).

- The loss of an experiment with sample size *n* given data *y* is

$$L(n, y, \theta) = |\theta - m_y| + 0.0008n$$

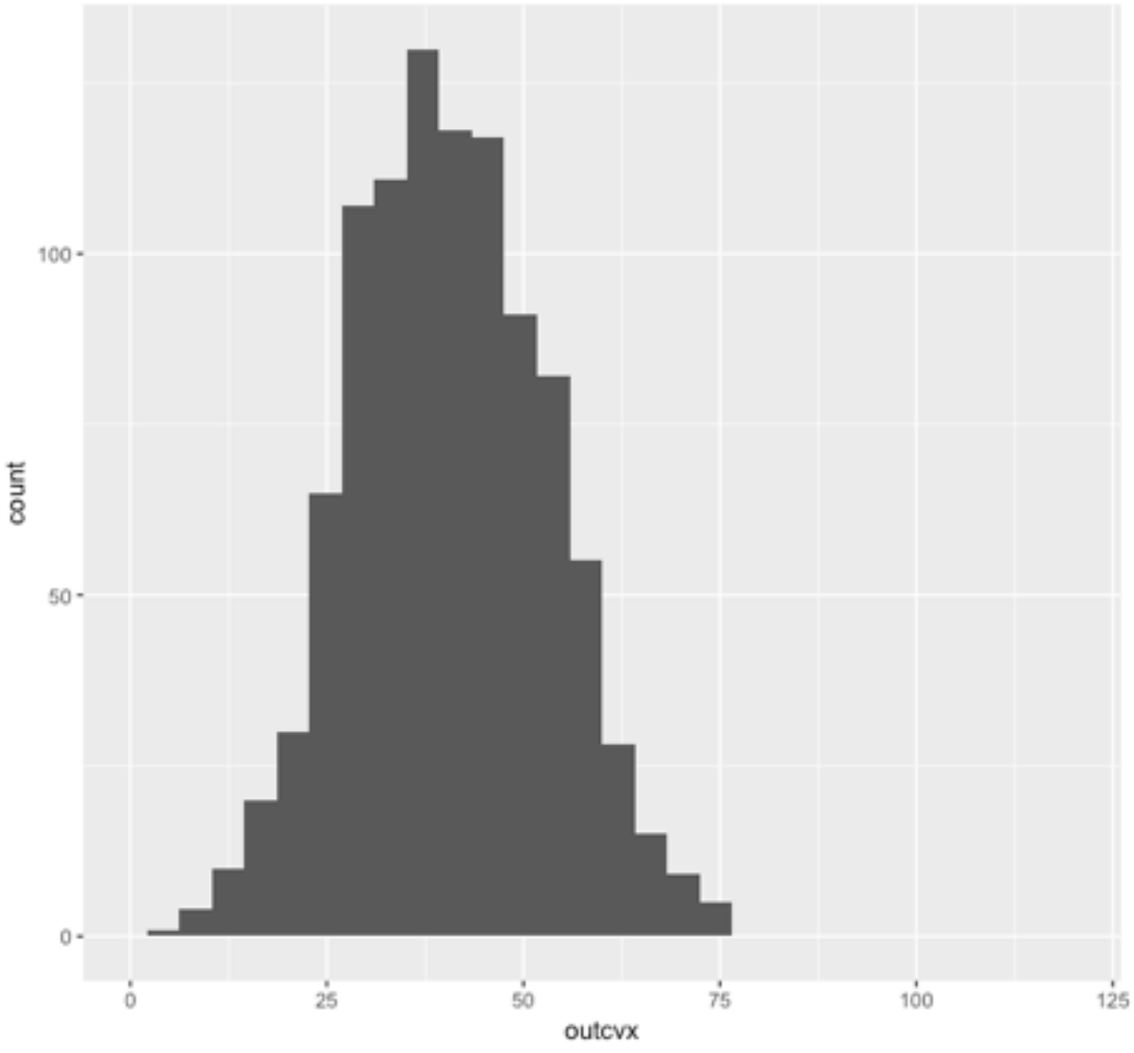
- The loss is estimated as the MC average of 100 simulations from the joint distribution of the data and $\theta$.

*This is Example 1 in Müller & Parmigiani (1995)*

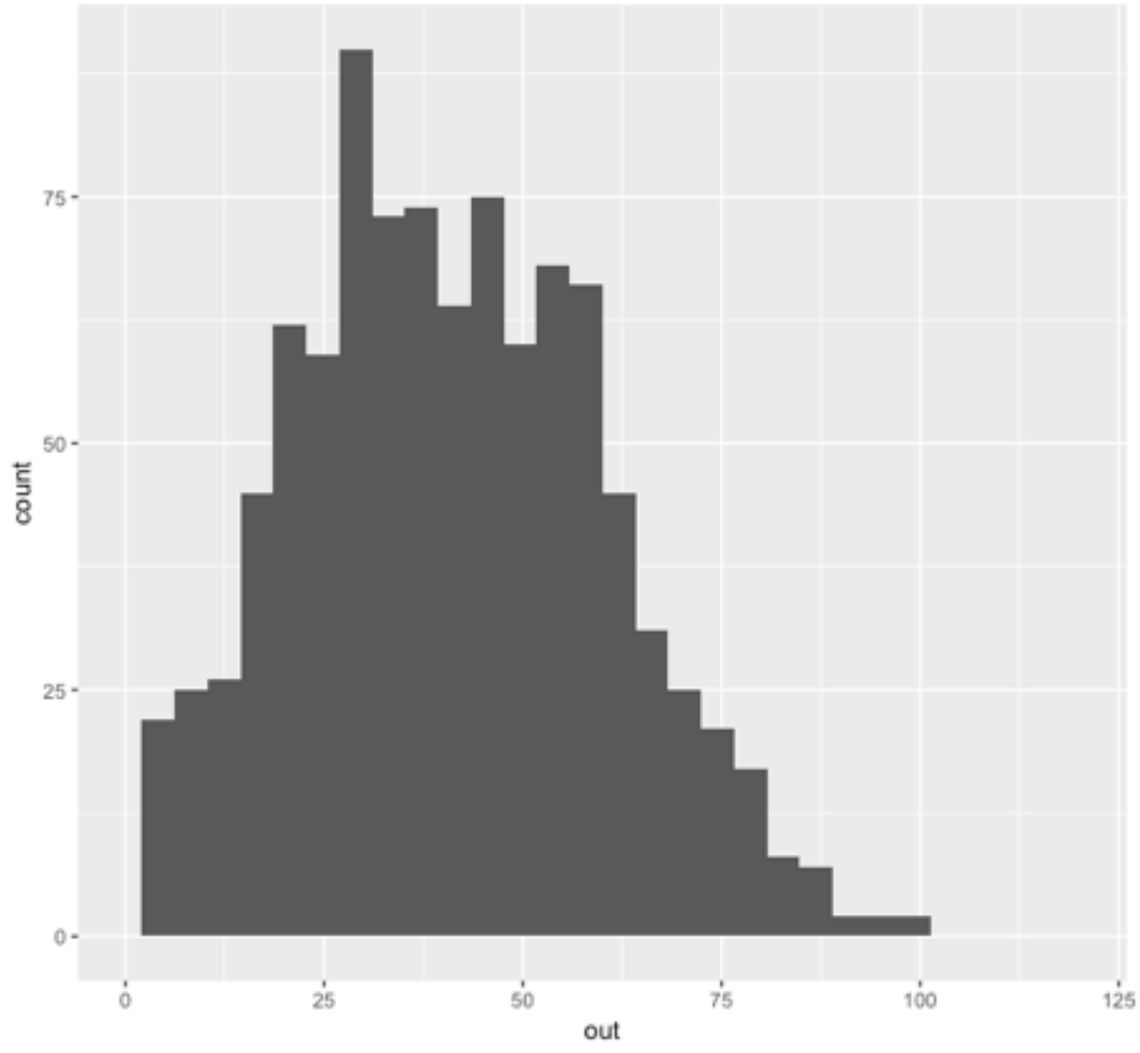


# Uncertainty on optimal sample size

Convex, 3 steps

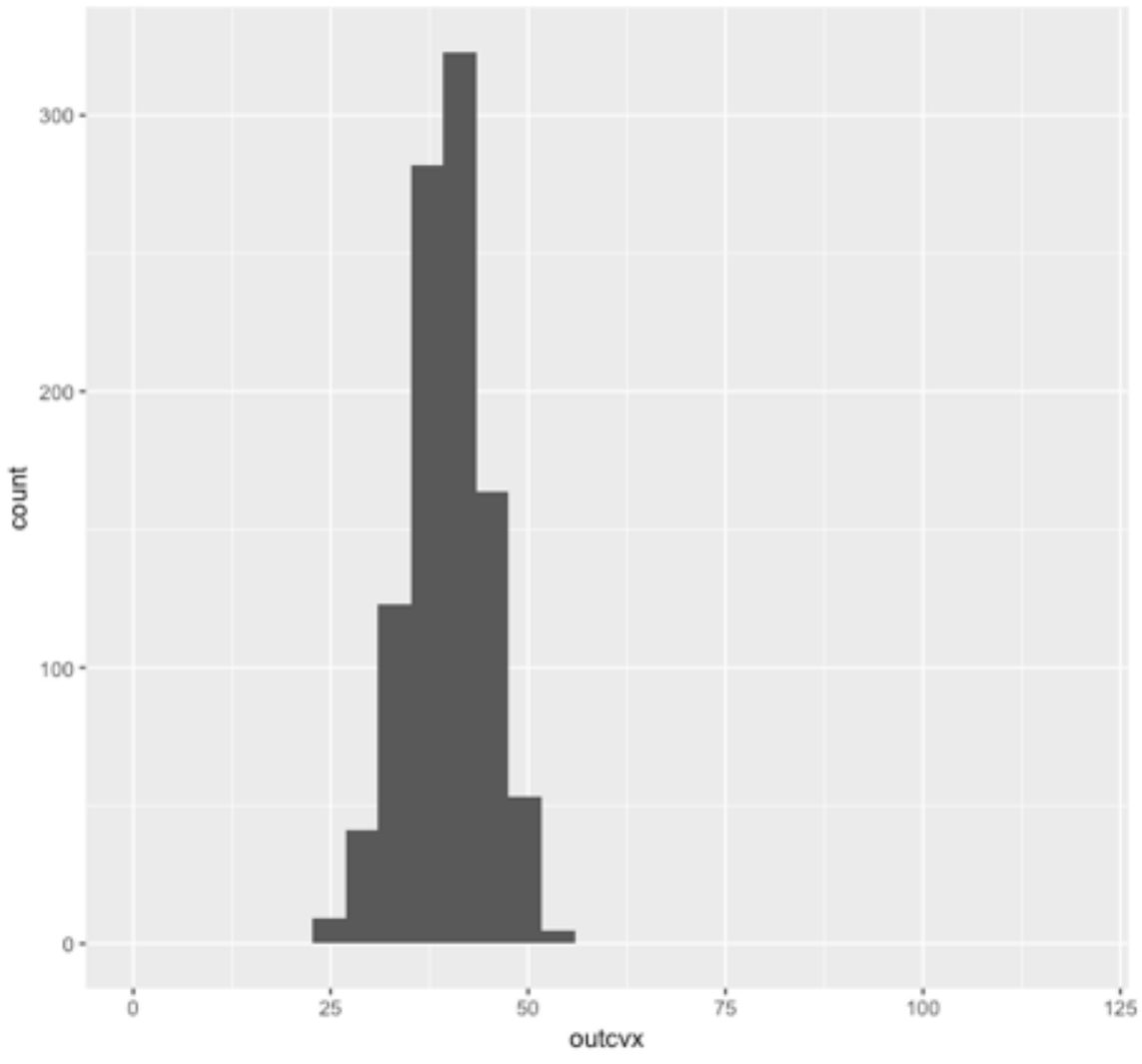



Unconstrained, 3 steps

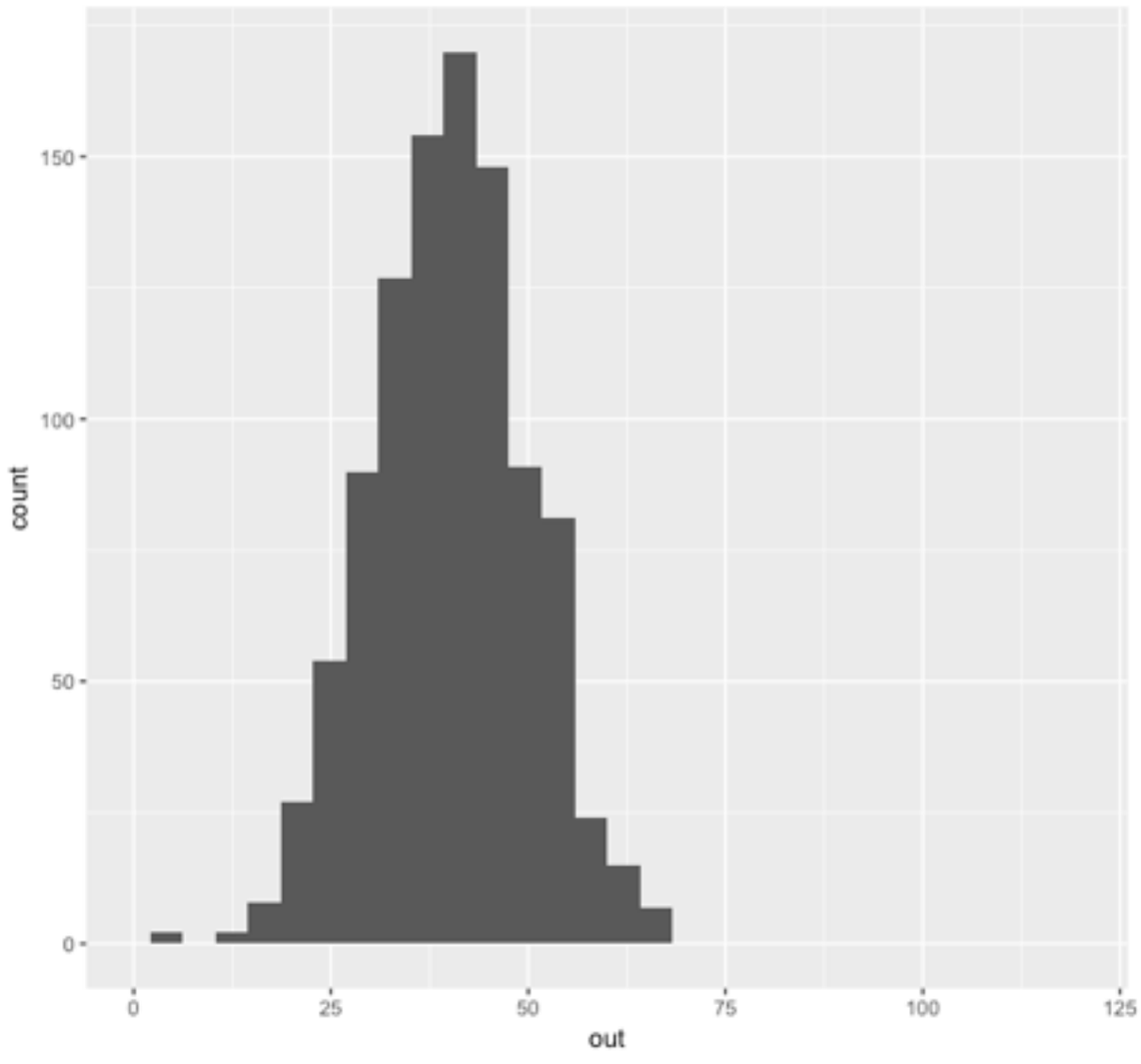


# Uncertainty on optimal sample size

Convex, 8 steps




Unconstrained, 8 steps



# Wrap-up

- When we run a numerical method, especially if we can't afford to run it "long enough," **we don't know what the right answer is**.

- It makes sense to quantify that uncertainty **probabilistically**.

- Bayesian Optimization uses GPs as surrogate models for objective functions.

- Prior information about the shape of the objective function is often available, so we should use it.

# References


Poincaré, H. (1896). *Calcul des probabilités*. Paris: Gauthier-Villars.

Diaconis, P. (1988). Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV, 1*, 163–175.

Hennig, P., Osborne, M. A., & Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 471(2179).*

Lindley, D. V. (1956), On a measure of information provided by an experiment. *Annals of Mathematical Statistics, 27 (4):* 986–1005.

Attolini, C. S. O., Peña, V., & Rossell, D. (2015). Designing alternative splicing RNA-seq studies. Beyond generic guidelines. *Bioinformatics, 31(22), 3631-3637.*

*Müller, P., & Parmigiani, G. (1995). O*ptimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association, 90(432),* 1322-1330.

Papoulis, A., & Pillai, S. U. (2002). *Probability, random variables, and stochastic processes*. McGraw-Hill Education.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for regression.* MIT Press.

Wang, X. & Berger, J.O. (2016). Estimating shape-constrained functions using Gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification, 4(1).* 1-25.

Z. I. Botev. (2016) The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).*


**Check out probabilistic-numerics.org/literature/ and bayesopt.com for more.**