An importance sampling procedure for estimating crop yield

HARVARD School of Engineering and Applied Sciences

J. Benjamin Cook

IACS Institute for Applied Computational Science
HARVARD SCHOOL OF ENGINEERING AND APPLIED SCIENCES

Introduction

Consider a corn field Ω and a yield function $f(x)$ that returns bushels per acre at any location in the field x. Total yield, in bushels can be evaluated by multiplying the area of the field, $|\Omega|$, by the integral, $I = \int_\Omega f(x)dx$. Assuming we can evaluate the function f at any location in the field, we can estimate the integral with $\hat{I}_{mc} = \frac{1}{N}\sum_i f(x_i)$, where the x_i are drawn uniformly from the area Ω. However, since yield is not distributed evenly throughout the field, this "vanilla" Monte Carlo approach results in unnecessarily high variance. Instead, it is possible to draw samples from a (possibly unnormalized) proposal distribution that is somehow "close" to f and then correct for the fact that the samples are no longer uniform. With importance sampling, the integral is estimated as:

$$\hat{I}_{is} = \int f(x)dx = \int g(x)\frac{f(x)}{g(x)}dx \approx \frac{1}{N}\sum_{x_i \sim g(.)} \frac{f(x_i)}{g(x_i)}$$

This approach assumes that $f(x)$ and $g(x)$ are normalized probability densities. Alternatively, we can use unnormalized functions if we correct for them as follows:

$$\hat{I}_{is} = \sum_i w_i f(x_i)$$

where $w_i = \frac{\tilde{w}_i}{\sum_i \tilde{w}_i}$ and $\tilde{w}_i = \frac{f(x_i)}{g(x_i)}$ [1].

Data

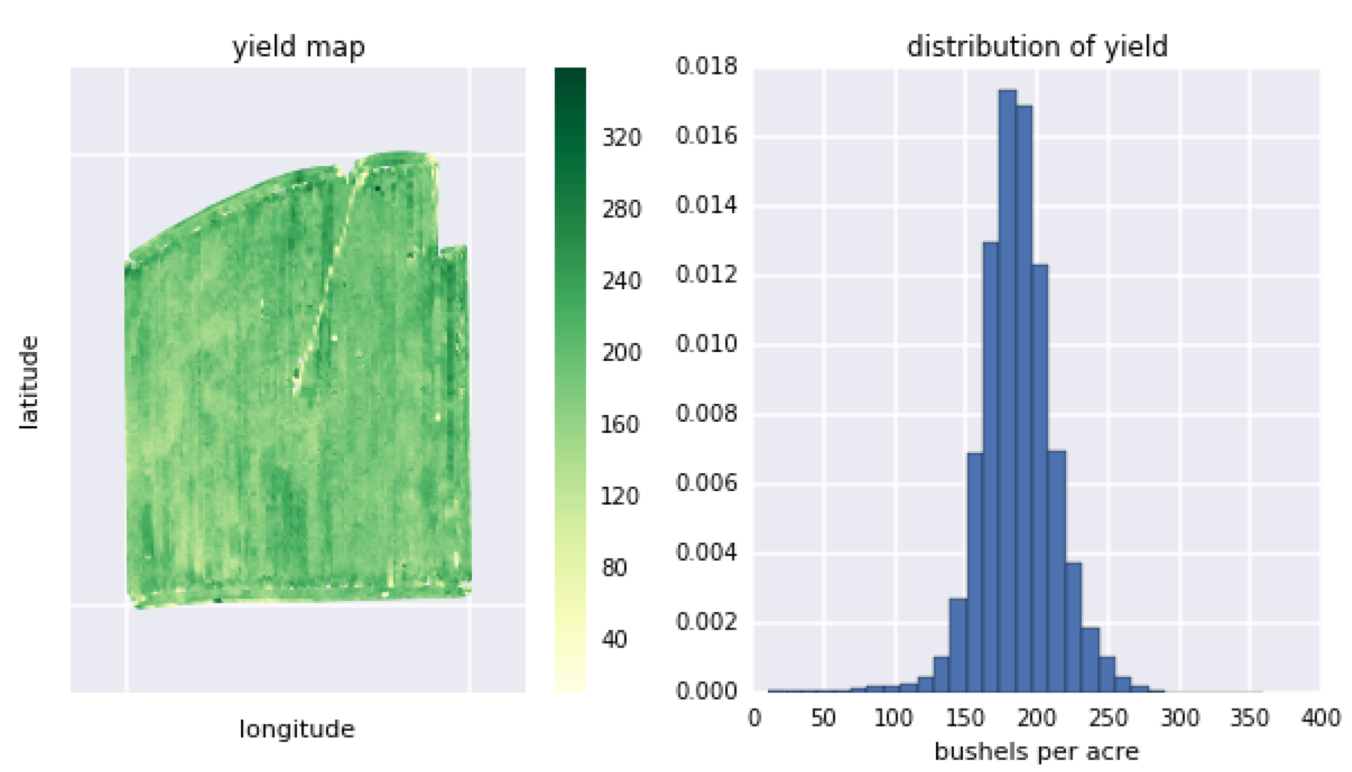


Figure 1: A yield map with the raw data and the distribution of yield at all measured locations

The data come from a corn field in Butler County, Nebraska that was planted and harvested in 2009. Measurements are logged by a grain yield monitor which is connected to sensors on the arms of the tractor as it harvests corn. The yield monitor records an estimate of yield in bushels per acre and geolocation approximately six times per second. This results in $N = 16,898$ points. Using rejection sampling, the number of acres is found to be approximately 63.7.

Method

The importance sampling procedure for estimating yield consists of four steps:

1. Collect a small number of samples from $f(.)$
2. Construct the proposal function $g(.)$
3. Sample from $g(.)$
4. Use the samples to estimate yield

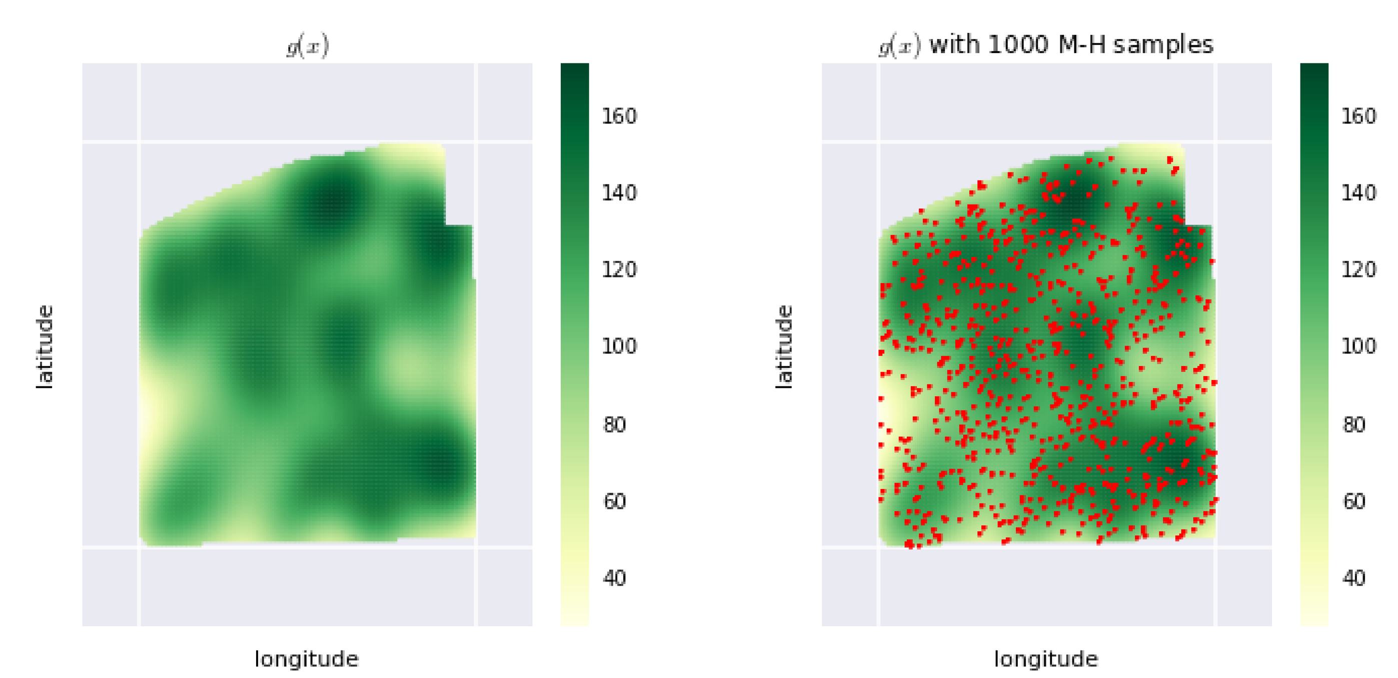
A Gaussian Process is a distribution over functions and, assuming the mean is set to zero, is fully specified by a covariance kernel, $K(x_i, x_j)$ [2]. One common form for the covariance function is a squared exponential:

$$\sigma^2 \exp\left(-\|x_i - x_j\|^2/\phi\right)$$

Assuming we know the hyper-parameters, σ^2 and ϕ, we now have a function that we can use for importance sampling:

$$g(x_{new}) = K(x_{new}, X)K^{-1}(X, X)y$$

which corresponds to the mean of the GP. Here, X is the vecotor of locations for the N_1 samples and x_{new} is the location of any arbitrary new point.



Mean of GP | MH samples

Figure 2: GP with Metropolis samples

Next, we can take $g(x)$ to be an unnormalized probability density and sample from it in order to find good candidate points for evaluating crop yield. I start with a simple Metropolis-Hastings sampler with symmetric proposals $q(x^*|x) \sim \mathcal{N}(x, \gamma)$, where γ is set to 0.002 [3].

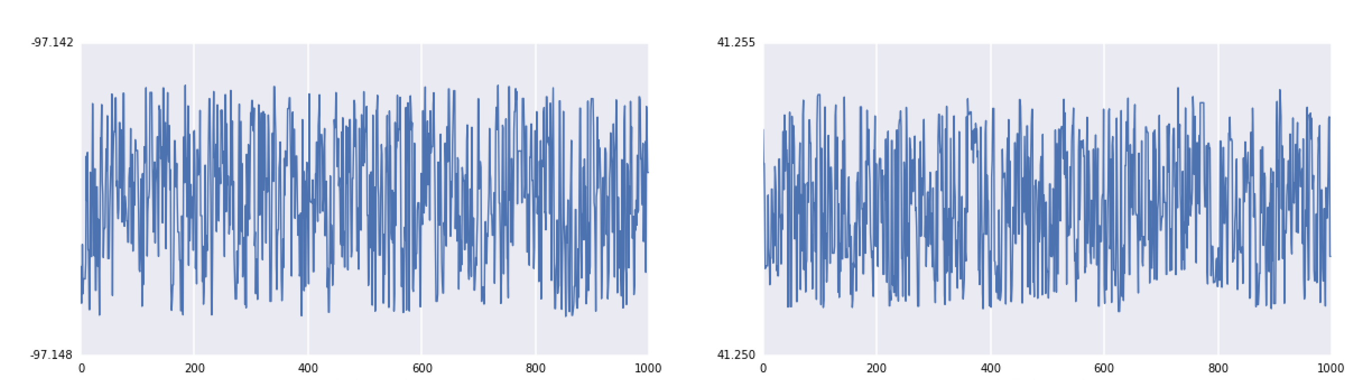


Figure 3: Trace plots of longitude and latitude samples from the M-H algorithm

Results

Unsurprisingly, as the number of M-H draws increases, the the variance of the estimate decreases. Table 1 show the average estimate and the standard deviation of those estimates for $N_{mc} = 100, 500$, and $1,000$. Each experiment was performed 100 times. The left panel of Figure 4 shows histograms of the three levels of M-H draws and the right panel shows a box plot of the estimates and variance. [H]

Table 1: Results

N_{mc}	Bushels	sd($\hat{I}_{is}$)
100	11,987	160
500	11,979	79
1,000	11,988	58

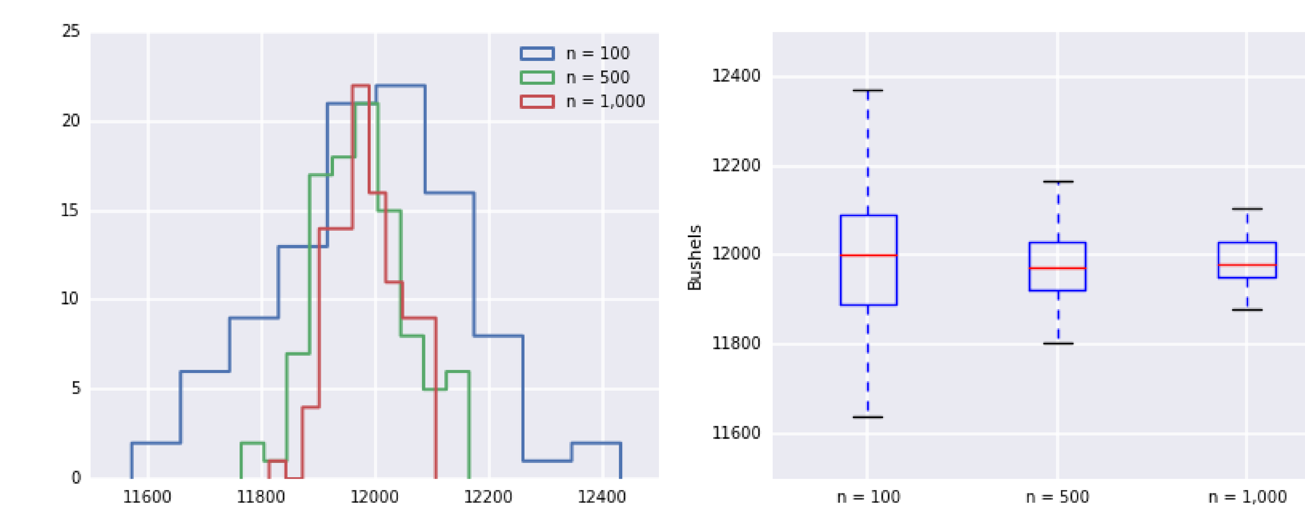


Figure 4: Distribution of estimates of $|\Omega|\hat{I}_{is}$ and variance of estimate as the number of Metropolis samples increases

In practice, how many M-H samples need to be drawn would have to be determined by the farmers and agronomists who use this procedure. Samples are relatively expensive in the sense that each one needs to be collected by hand. Fortunately, even with only $N_{mc} = 100$ samples, the standard deviation of 160 is less than 2 % of the number of bushels. This low variance makes the importance sampling method practical for estimating yield in real situations.

Conclusions

Stochastic optimization provides an important set of tools for estimating the crop yield in irregularly shaped fields. Fitting a GP to a few samples and then drawing several more importance samples from this proposal function decreases variance of estimated crop yield. This procedure can be used in a prediction setting with one additional step: by modeling how physical characteristics of the plant throughout the growing season drive yield at harvest.

References

[1] K. P. Murphy, *Machine Learning: a Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

[2] C. E. Rasmussen, *Gaussian processes for machine learning*. Citeseer, 2006.

[3] S. Chib and E. Greenberg, "Understanding the metropolis-hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.