Project: Propagating Monte Carlo error

Part I

Assignment

In many applications, a function $f(x)$ can only be estimated approximately, with some error. In this problem, you seek to estimate $I = \int_{-1}^{1} f(x)dx$ using $\hat{I} = \int_a^b \hat{f}(x)dx$, where $\hat{f}(x)$ is a least-squares polynomial estimate $\hat{f}(x) = \sum_{k=1}^{n_{fit}} \hat{c}_k x^k$: the $\hat{c}_k$ are chosen to minimize

$$-2\ln L = \sum_k \frac{(\hat{f}(x_k) - y_k)^2}{\sigma_k^2}$$

Your teams will be given data sets (x_k, y_k), where x_k are 11 uniformly distributed points from $-1, 1$; you can assume y_k are independent normally distributed random variables with variance $V(y_k) = \sigma_k^2 = 1$, with mean value $\langle y_k \rangle = f(x_k)$. You will be working with a **linear** fit, applied to functions that **may** be quadratic.

Preliminaries: Implement a code (or use existing routines) to find $\hat{c}_k$, the least-squares estimate, given any (x_k, y_k) and for a polynomial of arbitrary degree (i.e, `np.polyfit`) . Test your code with $f(x) = -2 + 3x$. Implement a code to generate random data with the desired statistical properties, given $f(x)$. Specifically, using the **fixed** values of x_k that I gave to you, generate standard normal random variables z_k and thus values y_k via

$$y_k = f(x_k) + \sigma z_k$$

In this expression, the $\hat{c}_k$ are the coefficients of this polynomial. For a linear fit, the relationship between $\hat{c}_k$ and $\hat{f}(x)$ are

$$\hat{f}(x) = \hat{c}_0 + \hat{c}_1 x \tag{1}$$

In this expression, x is a **parameter** (i.e., some to-be-determined constant) which is **not** random; note many of your answers will be expressions that depend on x.

Distribution of linear fit coefficients: The random variables $\hat{c}_k$ are linear combinations of normal random variables, and hence should each be normally distributed with some mean, some variance, and (critically for the next part) some *covariance*. Use your procedure to estimate the cumulative distribution of $\hat{c}_0, \hat{c}_1$, using (at least) 1000 random data sets, for the test function $f(x) = -2 + 3x$ described above. How does the mean of each coefficient depend on the choice for f? [Try $f(x) = 3$ and $f(x) = -2 + 3x$] Does the variance depend on the choice for f?

Covariance: Linear fit: The estimates $\hat{c}_k$ are generally correlated. Using a linear fit and synthetic data for (b), produce a scatterplot of $\hat{c}_0$ and $\hat{c}_1$ for 1000 samples. Argue the two variables are correlated. Estimate the variance of $\hat{c}_0, \hat{c}_1$, and the covariance $cov(\hat{c}_0, \hat{c}_1)$. Does the correlation coefficient depend on the choice for f?

Variance for $\hat{f}$: Linear fit: Using the results above, find an expression for $V(\hat{f}(x))$, the variance of the estimate $\hat{f}$ at x.

Confidence interval for $\hat{f}(x)$: Linear fit: Using the correlation coefficient provided above, assumed *known exactly*, plot $\hat{f}(x)$ and a 90% confidence interval $\hat{f}(x) \pm z_{0.05}\sqrt{V(\hat{f}(x))}$, where $\hat{f}(x)$ is estimated using the **real** data provided in this assignment.

Integral and integral error: Linear fit: Find an expression for $\hat{I}$ and the variance of $\hat{I}$ in terms of your results above. Verify this expression using your synthetic data. Provide a 90% confidence interval estimate for I based on the real data.

Extra credit: +100%: Repeat the steps above for the quadratic case.

Extra credit: +150%: [If you know linear algebra] Assume normal iid errors n_k with zero mean and unit variance, and assume y is determined from n and the constants c_α, x via $y_k = \sum_\alpha F_\alpha(x_k)c_\alpha + n_k$. For convenience, organize $F_\alpha(x_k)$ into a matrix $F_{k\alpha}$ and c into a row vector.

1. Show that least-squares regression implies that for the best-fitting c_*, $F^T(y - Fc_*) = 0$.
2. Assuming $F^T F$ is invertible, find the least-squares estimator c_* in terms of y and F
3. Show c_* is an unbiased estimator for c: $\langle c_* \rangle = c$.
4. Find the covariance matrix of c_* in terms of the covariance matrix of y and F.

Using this framework, find an expression for the covariance matrix for $\hat{c}$. Find an expression for the confidence interval for f using y_k, F, and $F^T F$.