Approximated PCA

Iteration 3

Rodrigo Arias

March 23, 2017

Proposed hypothesis

Let ϵ_b be the error in the result with width b, then the following hypothesis can be proposed:

- The mean of $\log_2(\epsilon_b) \approx -b$.
- The standard deviation ≈ 2.

Exploratory data analysis

The technique was proposed by John Tukey to staticians, to explore new data. The method works in three main steps:

1. Inspect visually
2. Propose hypothesis
3. Test the proposed hypothesis

Definition of X_b

- Considering the experiments performed with the bit-width b.
- With ϵ_b being the measure of the rounding error, let $X_b = \log_2(\epsilon_b)$ be a random variable.
- Then, after k runs, we obtain the sample means $\overline{X}_b$ for all values tested of b.

Plot of the sample mean error $\overline{X}_b$

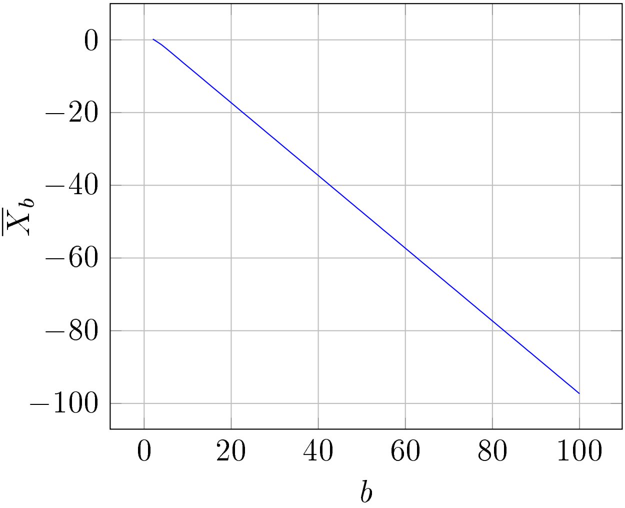


Figure: Plot of $\overline{X}_b$ as the number of bits b increases.

Difference with b

It seems that $\overline{X}_b$ is almost $-b$. To see if there is any difference, we can plot $b + \overline{X}_b$, and check if the mean of the result is 0.

Plot of the difference

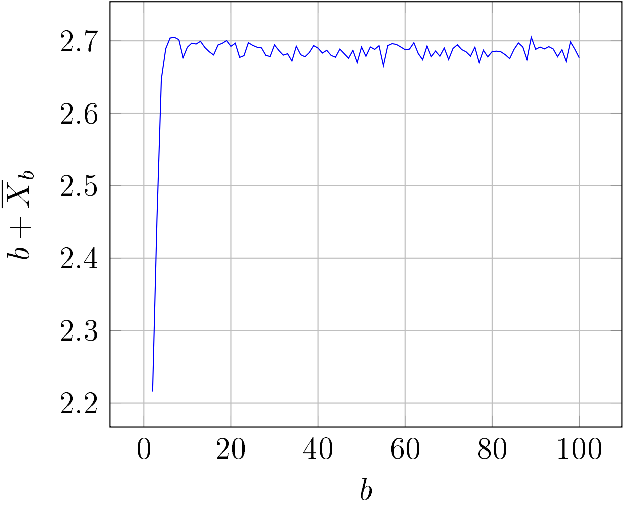


Figure: Plot of $b+\overline{X}_b$ as the number of bits b increases.

Deviation on lower bits

There can be seen that when $b < 5$, the error deviates from the mean. Lets ignore those values as outliers. And check again the plot.

Plot of the difference

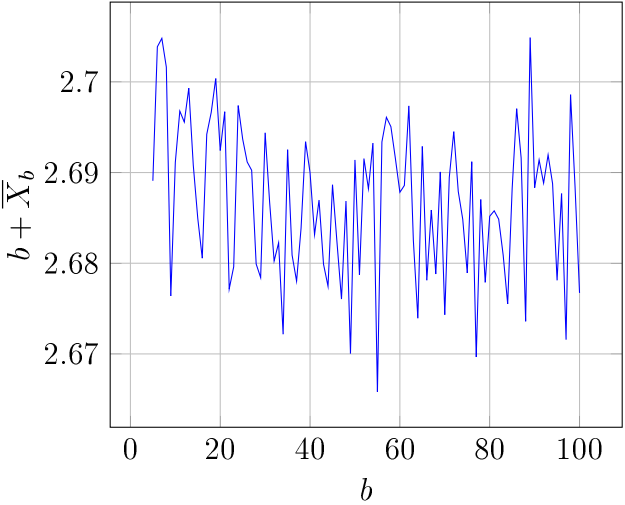


Figure: Plot of $b + \overline{X}_b$ as the number of bits b increases.

Histogram of the difference

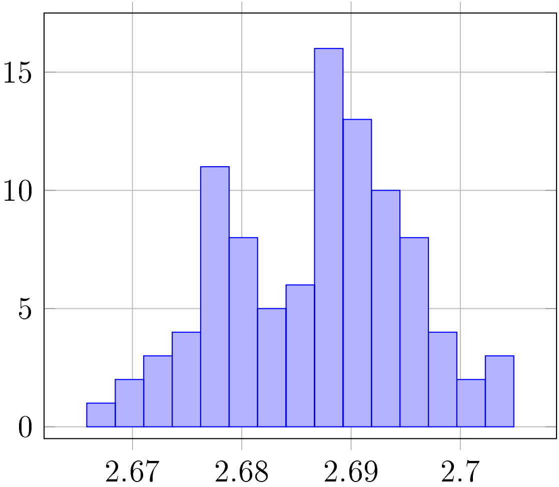


Figure: Histogram of the difference $\bar{b} + X_b$

Distribution of the error

Now the difference seems to be a random variable, with a fixed mean μ independent of b. Let $Y = b + \overline{X}_b$. Then,

$$\overline{X}_b = -b + Y$$

By the central limit theorem, with probability $1 - \alpha$, the mean μ will lie in the region:

$$\overline{Y} \pm z_{\alpha/2} S/\sqrt{n}$$

The mean μ is in 2.68691 ± 0.00190 with 95% of confidence.

Plot of the difference

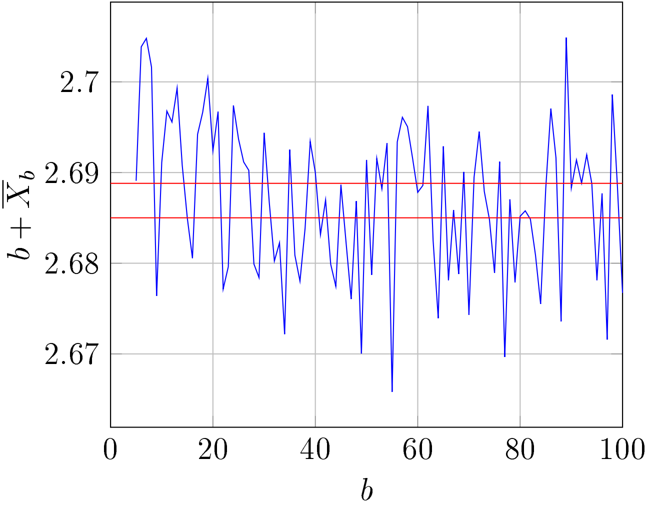


Figure: The mean μ is in the marked region with 95% confidence.

Test of independence

- To determine if the error is dependent of the bit-width b, a correlation test can be performed.
- The null hypothesis is that Y is dependent of b. The correlation coefficient is -0.222 and the p-value 0.02969.
- So we can reject the null hypothesis, and assume that they are independent.

Conclusions about the mean $\overline{X}_b$

- $\overline{X}_b$ is proportional to b. Can be described as $\overline{X}_b = -b + Y$.
- The random variable Y is independent of b.
- There is a strange behavior when $b < 5$, where $\overline{X}_b$ don't follow the expected value. The cause of this behavior is yet unknown.
- More experiments are needed, to test if X_b depends on the input size N.

Observation of the standard deviation

The sample standard deviation S can be observed for each bit-width b.

Let S_b be the sample s.d. of X_b in the k runs.

Note that the real standard deviation σ_b is unknown, but S_b is an unbiased estimator of σ_b

Plot of the standard deviation

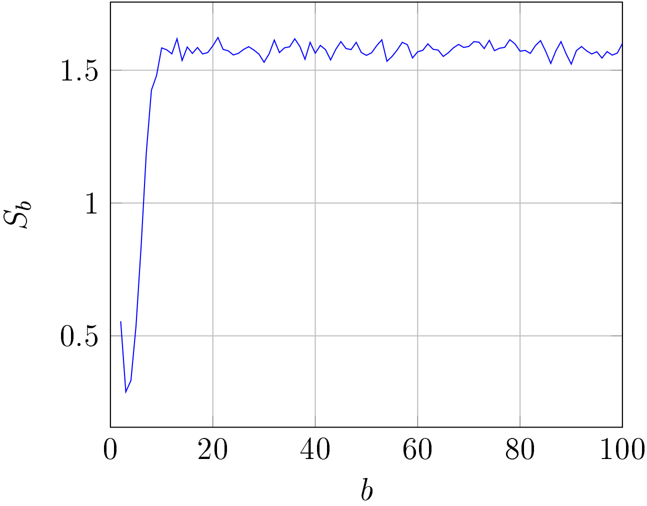


Figure: The sample standard deviation S_b as the bit-width b grows.

Deviation on lower bits

Again, there is a deviation from the mean on the lower bits, now when $b < 12$. Those values will be ignored as outliers. Also, the mean of the standard deviation seems to be near 0.9 not 2.

Plot of the standard deviation

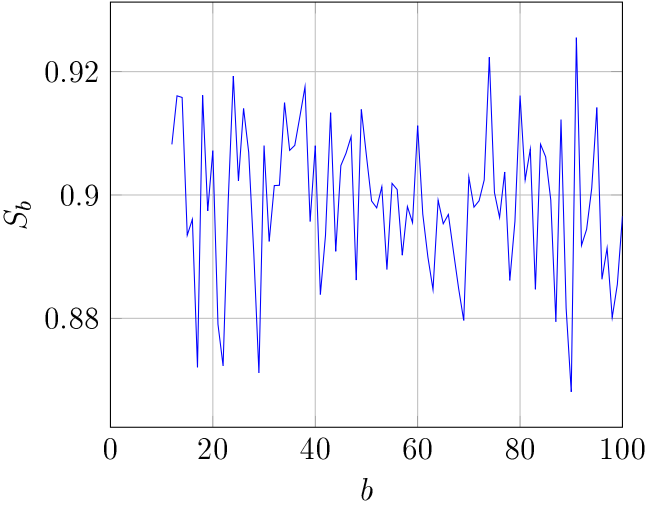


Figure: The sample standard deviation S_b as the bit-width b grows.

Histogram of S_b

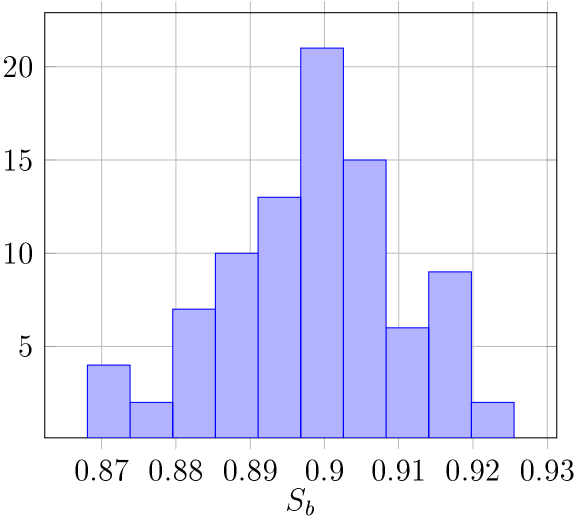


Figure: Histogram of the standard deviation S_b

Tests of S_b

First, it can be tested if S_b is independent of b. The correlation coefficient is -0.158, but there is not enough confidence to reject the hypothesis of dependent variables.

The mean deviation is 0.8989.

Conclusions about the S_b are not yet complete.

Variable precision in Householder

A new experiment is now performed keeping different bit-widths by variable in the Householder algorithm.

At each run, only one variable is tested, while the others remain with high precision.

The results show how each variable affects the overall error.

Plot by variable

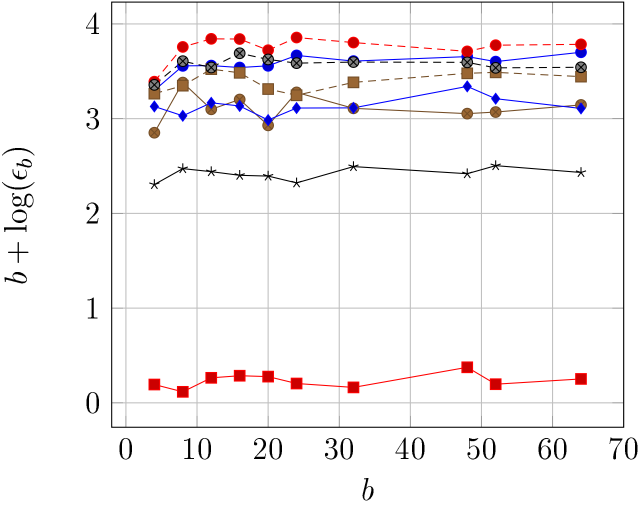
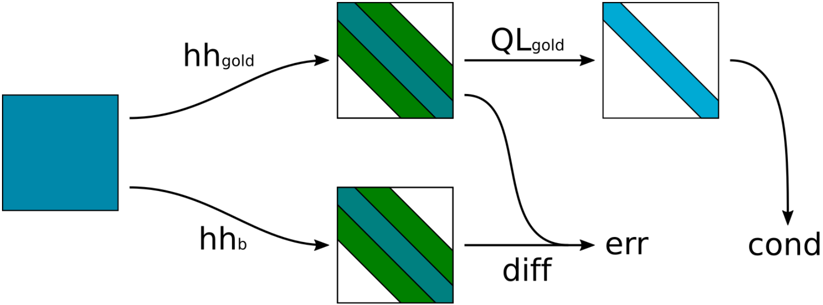


Figure: Different error mean for each variable.

Condition number


hh$_{gold}$
QL$_{gold}$
hh$_b$
err
cond
diff