1. **What happens when the prior mean is initialized to the wrong value? Correct value?**

   The figure below is when I run my script for a mean and variance of: 0.43 and 0.9 and the prior values are intialized to: 0.87 and 2

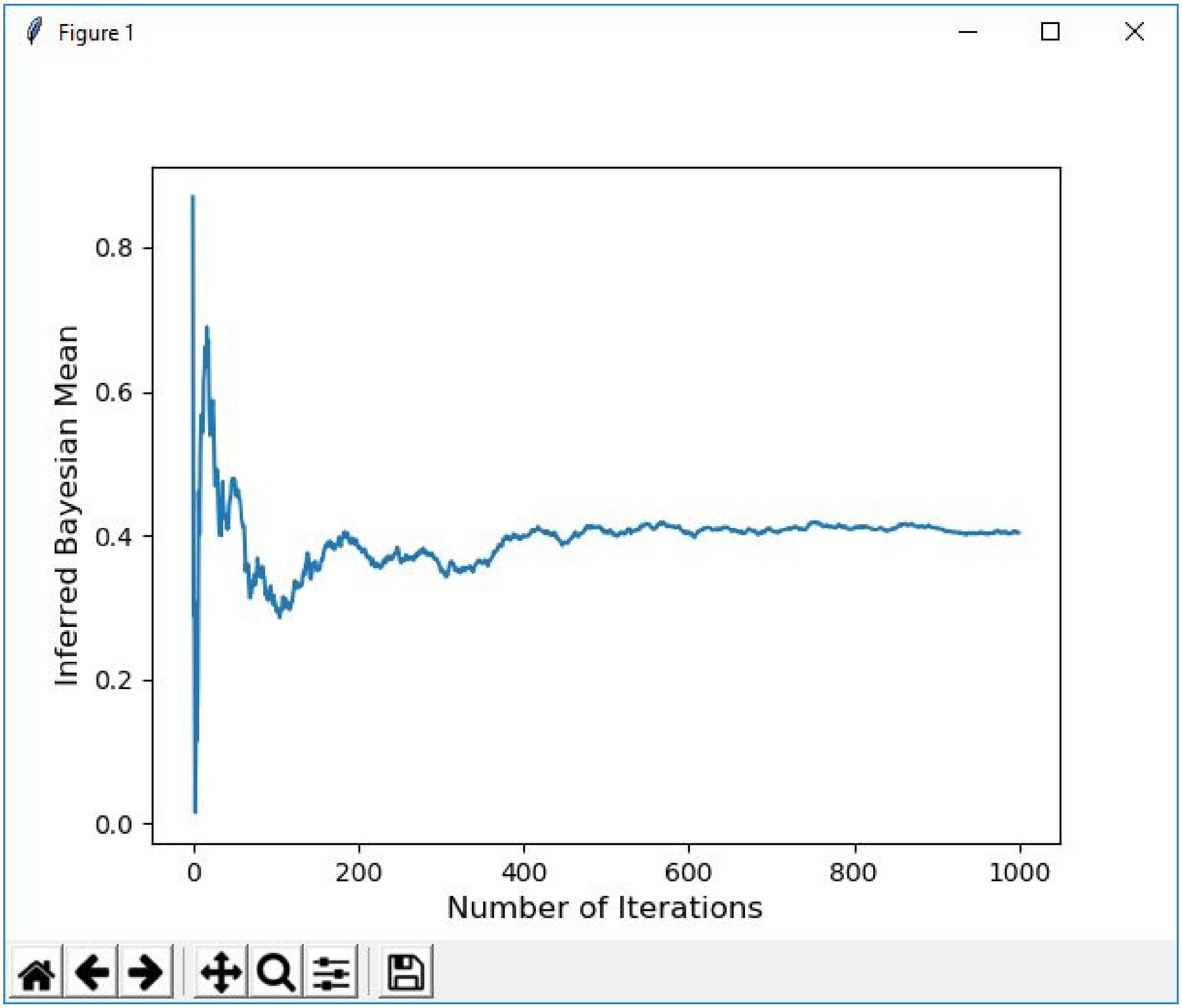


Figure 1-1: The unstable oscillations are due to the incorrect initialization values, a guess would be to say that because my guessed mean is greater than a factor of 2 that the oscillations can reach that upper bound and with enough iterations slowly converge to 0.43

The figure below is when I run my script for a mean and variance of: 0.43 and 0.9 and the prior values are intialized to: 0.43 and 2

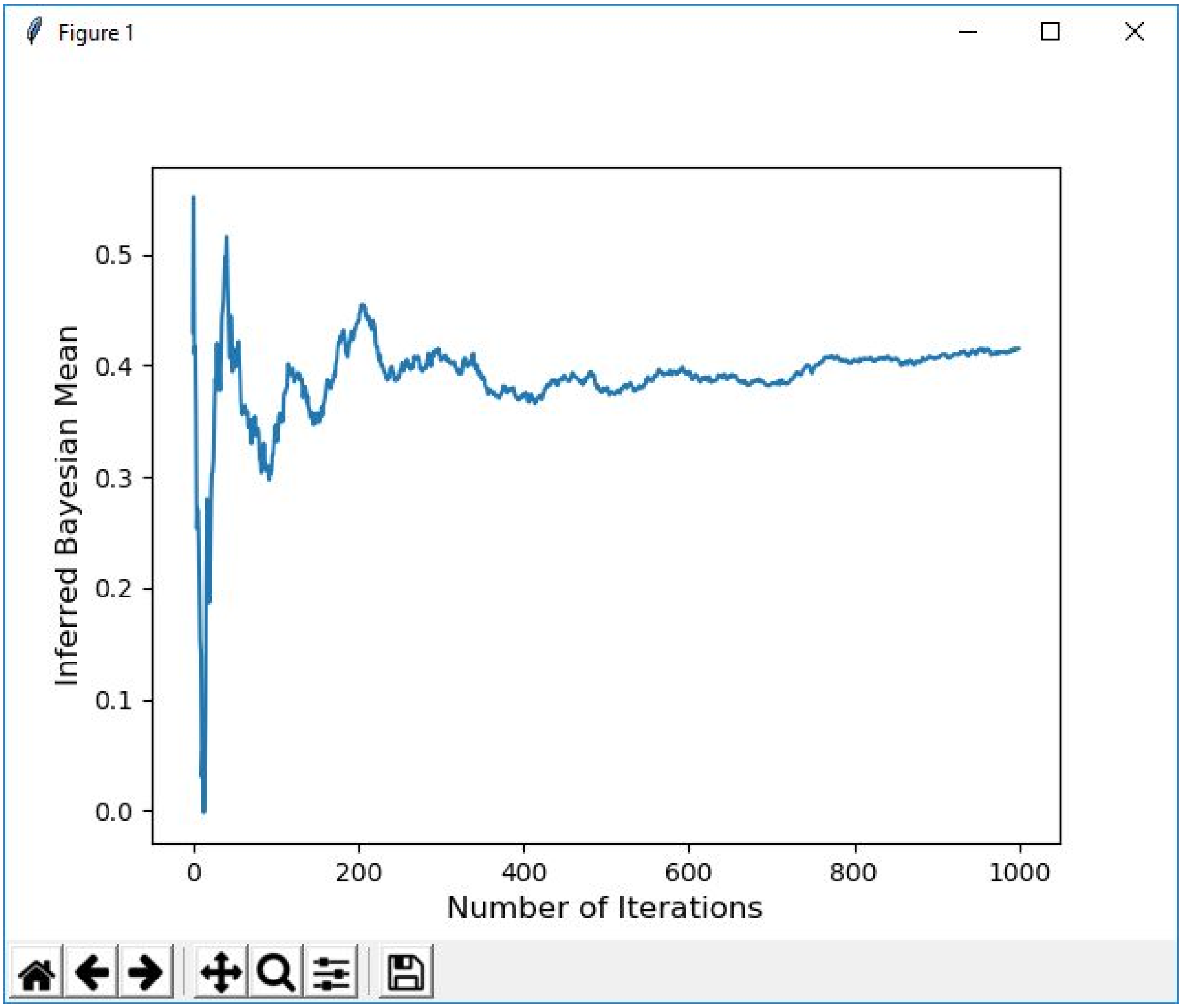
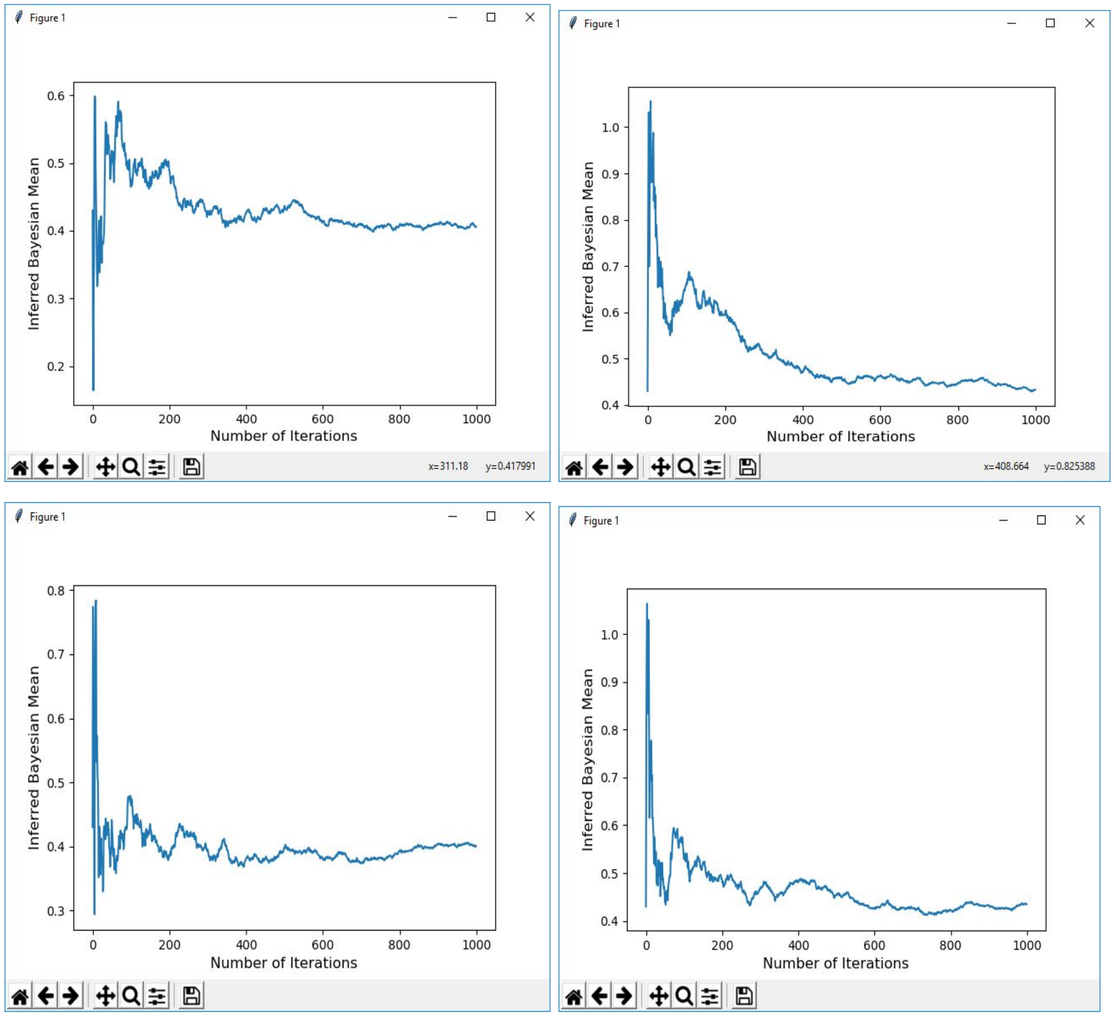


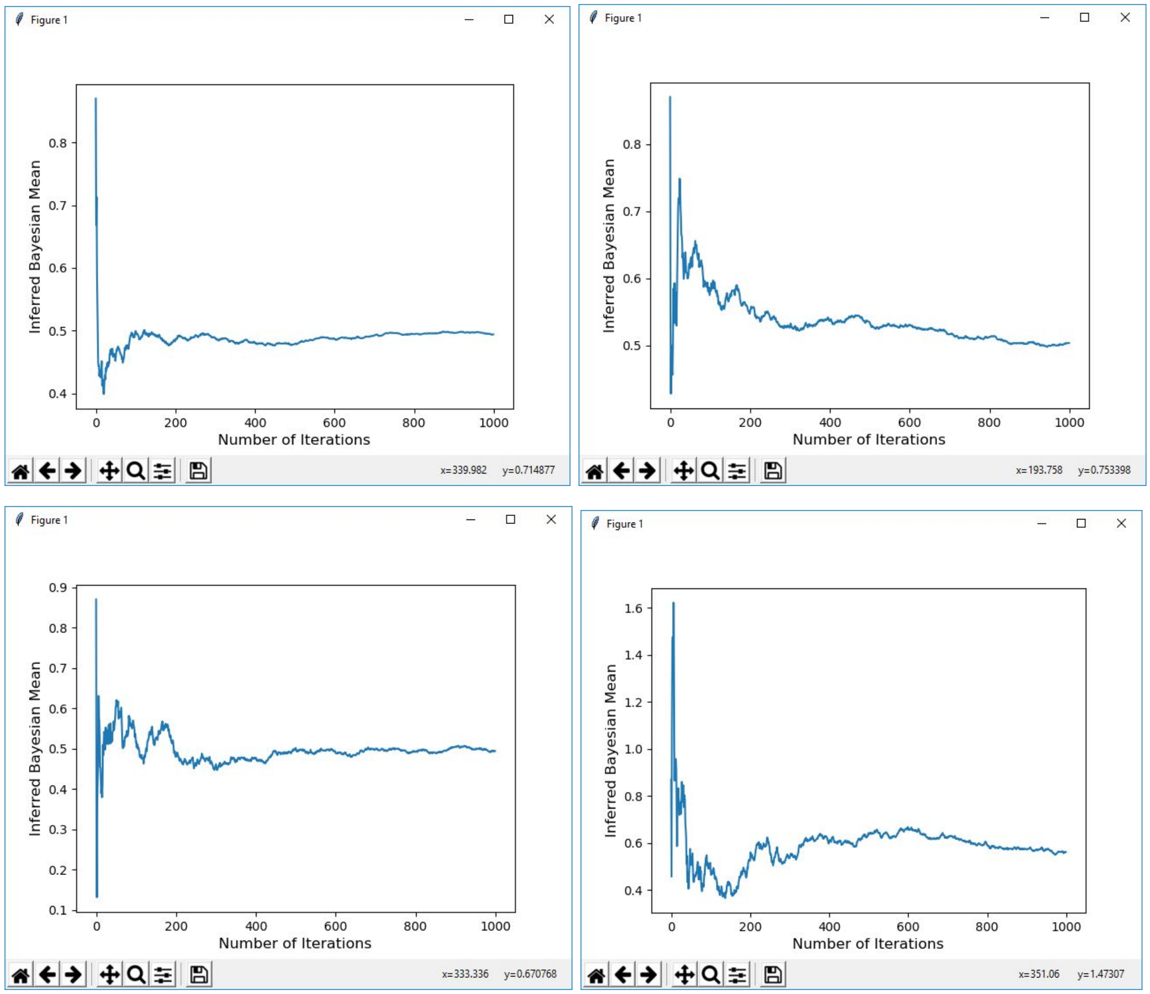
Figure 1-2: Now the oscillations peak at a lesser value and values close to the ML appear quicker compared to Figure 1-1.

2. **What happens as you vary the prior variance from small to large?**



Figures 2-1 - 2-4: As prior variance increases: 0.5,1,2,10
(from left to right and top to bottom) the certainty of the approximation decreases and the early estimations are very inconsistent as noted by the purely vertical lines which show that the successive values chosen are vastly different.

3. **What happens when the likelihood variance is varied from small to large?**



Figures 3-1 - 3-4: As likelihood variance increases: 0.25,0.5,1,2

(from left to right and top to bottom) When the original data has too much variance and the prior estimations are kept constant, the convergence rate to the mean takes more iterations of the program because the probabilty of choosing the correct estimate is much less likely to occur within a certain amount of execution of the program.

4. **How do the inital values of the prior mean , prior variance, and likelihood variance interact to effect the final estimate of the mean?**

The parameters mentioned all contribute to calculating the posterior estimate which is a consequence of Bayes Theorem. Using it we can restate it as: posterior is very closely approximated by taking the likelihood of an event (in this case, a defined random variable with mean and variance) and multiplying it by a prior distribution (adding subjective beliefs and assumptions about the random variable). In our case, the prior and posterior distributions are the same type of distributions so there is a conjugate distribution that we used when the random variable is normally distributed.
As an outcome of Bayes rule we can say that when the prior is not descriptive, the posterior will be mostly modeled from data that is available.
When the opposite is true, the posterior will become more ‘true’ the more data it has in order to ‘merge’ and overcome biases.

Discussion: MLE vs MAP concepts and insight

Together these provide insight into estimating the value of a variable notice how this relates to a single estimated number It does not extend into approximating full distributions. In class, we decided to work in a logarithm space because optimazation is easier to calculate for a log that is monotonic. When the result is compared to Bayes Theorem, we found that they are very similar and the only thing missing was a scalar factor in the denominator. So we could relate the theorem as: Probability of event X given Y = Probablity of event Y given X multiplied by Probablity of event X. Looking back to the Maximum likelihood estimation formula it only differes by a factor of: Probabilty of X, which in this case describes the prior!
Recall that the optimization did not include the denomitor part of Bayes Rule because it was a scalar value. So an important conclusion can be drawn that MLE is only different from MAP when the prior value is not a constant. For example, in class we needed to account for a prior distribution for a binomially distributed event. That is to say that because the prior can take on multiple values, the probabilty must be accounted for in some way and that is why we use MAP.

Program runs:

NOTE* first two runs are (mean = 0.5, variance = 1)

Prior mean = 0.8 and prior variance = 1

Last run is:

(mean = 0.69, variance = 2)

Prior mean = 0.3 and prior variance = 1

```
Max number of data points 1000
MLE:0.33352380783196645
MAP:0.5667619039159832
Press enter to continue...

MLE:0.35448865017898573
MAP:0.5029924334526572
Press enter to continue...

MLE:0.49933288257857594
MAP:0.574499661933932
Press enter to continue...

MLE:0.5017802912915623
MAP:0.56142423303325
Press enter to continue...

MLE:0.7302329762619595
MAP:0.7418608135516329
Press enter to continue...

MLE:0.5710997100552441
MAP:0.6037997514759235
Press enter to continue...

MLE:0.611717387442069
MAP:0.6352527140118104
Press enter to continue...

MLE:0.625693866271728
MAP:0.6450612144637581
Press enter to continue...
```

Figure E1: showing the speed of convergence in a 8 iterations.

```
Max number of data points 1000
MLE:-0.6046483021340843
MAP:0.09767584893295789
Press enter to continue...

MLE:-0.17920946590502768
MAP:0.14719368939664823
Press enter to continue...

MLE:0.18091295263163987
MAP:0.33568471447372994
Press enter to continue...

MLE:0.22024952375800888
MAP:0.33619961900640716
Press enter to continue...

MLE:0.5596060534430791
MAP:0.599671711202566
Press enter to continue...

MLE:0.6908844003357268
MAP:0.7064723431449089
Press enter to continue...
```

Figure E2: A demonstration with negative values.

```
MLE:0.8597557180364318
MAP:0.8149752605935177
Press enter to continue...

MLE:0.8105461757522132
MAP:0.7705033384383144
Press enter to continue...

MLE:0.72548126128001
MAP:0.6927519334892402
Press enter to continue...

MLE:0.7612437842461258
MAP:0.7264329326049093
Press enter to continue...

MLE:0.7128276574391836
MAP:0.6822478309622074
Press enter to continue...

MLE:0.6856640062696815
MAP:0.6576157149046139
Press enter to continue...

MLE:0.6678018162939955
MAP:0.6415302579872817
Press enter to continue...

MLE:0.6962977976104872
MAP:0.668487425848348
Press enter to continue...
```

Figure E3: after 29 iterations the mean is approached.

Ishmael Contreras
Assignment 4

Prove: $\mu_n = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}$ where $\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

Let the likelihood be $p(D|\mu,\sigma^2) = \prod_{i=1}^{n} p(x_i|\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$

where D = Data vector. $\{x_1, x_2 \ldots x_n\}$

Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{N} x_i$ be the mean (empirical)

$\sigma^2 = \frac{1}{n}\sum_{i=1}^{N}(x_i - \bar{x})$ be the variance (empirical)

rewrite the exponent:

$$\sum_i (x_i-\mu)^2 = \sum_i [(x_i-\bar{x}) - (\mu-\bar{x})]^2$$
$$= \sum_i (x_i - \bar{x})^2 + \sum_i (\bar{x}-\mu)^2 - 2\sum_i (x_i-\bar{x})(\mu-\bar{x})$$
$$= ns^2 + n(\bar{x}-\mu)^2$$

note * $\sum_i (x_i-\bar{x})(\mu-\bar{x}) = (\mu-\bar{x})\left(\left(\sum_i x_i\right) - n\bar{x}\right) = (\mu-\bar{x})(n\bar{x}-n\bar{x}) = 0$

Hence, $p(D|\mu,\sigma^2) = \frac{1}{2\pi^{n/2}}\frac{1}{\sigma^n}\exp\left(-\frac{1}{2\sigma^2}[ns^2 + n(\bar{x}-\mu)^2]\right)$

$$\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right)\exp\left(-\frac{ns^2}{2\sigma^2}\right)$$

with $\sigma^2 = c$

$$p(D|\mu,\sigma^2) = \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right) \propto \mathcal{N}\left(\bar{x}|\mu, \frac{\sigma^2}{n}\right)$$

the natural conjugate prior.

$$p(\mu) \not\propto \; \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \propto \mathcal{N}(\mu \mid \mu_0, \sigma_0^2)$$

the posterior

$$p(\mu|D) \propto p(D|\mu,\sigma)\, p(\mu|\mu_0,\sigma_0^2)$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right]\cdot \exp\left[-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$

$$= \exp\left[\frac{-1}{2\sigma_0^2}\sum_{i=1} x_i^2+\mu^2-2x_i\mu) + \frac{-1}{2\sigma_0^2}(\mu^2+\mu_0^2-2\mu_0\mu)\right]$$

$$p(\mu|D) \propto \exp\left[-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)+\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i x_i}{\sigma_0^2}\right)-\left(\frac{\mu_0^2}{2\sigma_0^2}+\frac{\sum_i x_i^2}{2\sigma_0^2}\right)\right.$$

$$= \exp\left[-\frac{1}{2\sigma_n^2}(\mu^2-2\mu\mu_n+\mu_n^2) = \exp\left(-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right]\right.$$

by matching μ^2 and σ_n^2 we arrive at:

* product of two gaussians is a gaussian

$$\frac{-\mu^2}{2\sigma_n^2} = \frac{-\mu^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)$$

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2}+\frac{n}{\sigma^2} \quad \checkmark$$

$$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2} = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}$$

matching powers of μ:

$$\frac{-2\mu\mu_n}{-2\sigma_n^2} = \mu\left(\frac{\sum_{i=1}^{N} x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{\sum_{i=1}^{n} x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2} \Rightarrow \frac{\sigma_0^2 n\bar{x}+\sigma^2\mu_0}{\sigma^2\sigma_0^2}$$

Hence, $\mu_n = \frac{\sigma_0^2}{n\sigma_0^2+\sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\bar{x} = \sigma_n^2\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\bar{x}}{\sigma^2}\right)$ /