1 Gaussian posterior credible interval

(Source: DeGroot)
Let \( X \sim \mathcal{N}(\mu, \sigma^2 = 4) \) where \( \mu \) is unknown but has prior \( \mu \sim \mathcal{N}(\mu_0, \sigma_0^2 = 9) \). The posterior after seeing \( n \) samples is \( \mu \sim \mathcal{N}(\mu_n, \sigma_n^2) \). (This is called a credible interval, and is the Bayesian analog of a confidence interval.) How big does \( n \) have to be to ensure
\[
p(\ell \leq \mu_n \leq u | D) \geq 0.95
\]
where \((\ell, u)\) is an interval (centered on \( \mu_n \)) of width 1 and \( D \) is the data. Hint: recall that 95% of the probability mass of a Gaussian is within \( \pm 1.96\sigma \) of the mean.

2 MAP estimation for 1D Gaussians

(Source: Jaakkola)
Consider samples \( x_1, \ldots, x_n \) from a Gaussian random variable with known variance \( \sigma^2 \) and unknown mean \( \mu \). We further assume a prior distribution (also Gaussian) over the mean, \( \mu \sim \mathcal{N}(m, s^2) \), with fixed mean \( m \) and fixed variance \( s^2 \). Thus the only unknown is \( \mu \).

1. Calculate the MAP estimate \( \hat{\mu}_{MAP} \). You can state the result without proof (see Section ??). Alternatively, with a lot more work, you can compute derivatives of the log posterior, set to zero and solve.

2. Show that as the number of samples \( n \) increase, the MAP estimate converges to the maximum likelihood estimate.

3. Suppose \( n \) is small and fixed. What does the MAP estimator converge to if we increase the prior variance \( s^2 \)?

4. Suppose \( n \) is small and fixed. What does the MAP estimator converge to if we decrease the prior variance \( s^2 \)?

3 Language modeling with the Dirichlet-multinomial model

Consider the following children’s nursery rhyme:
mary had a little lamb, little lamb, little lamb,
mary had a little lamb, its fleece as white as snow

Let us convert this (after removing punctuation marks like commas) to a string of integers using the mapping
mary = 1, had = 2, a = 3, little = 4, lamb = 5, its = 6, fleece = 7,
as = 8, white = 9, snow = 10

Thus we get
\[
D = (1, 2, 3, 4, 5, 4, 5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 8, 10)
\]
where \( D = (X_1, \ldots, X_{20}) \) is the data and \( X_i \in \{1, \ldots, 10\} \) is the identity of the \( i \)'th word. (Thus the vocabulary has size \( K = 10 \).) Assume \( X_i \sim \text{Discrete}(\theta) \) are iid random variables, so \( p(X_i = j | \theta) = \theta_j \). Let \( p(\theta) = \text{Dir}(\theta | \alpha_1, \ldots, \alpha_{10}) \), where \( \alpha_j = 1 \) for all \( j \).
1. What is the posterior predictive distribution \( p(\tilde{X}|D) \)? (This should be a histogram of 10 numbers). (Here \( \tilde{X} \) represents a new word sampled from the distribution.)

2. What is the most probable next word in the sentence, \( \arg \max_j p(\tilde{X} = j|D) \)? (There may be more than one answer.)

3. How might this language model be improved? (Give a brief (2-3 sentence) description of any ideas you have.)

4. **MAP estimation for the Bernoulli with non-conjugate priors**

   (Source: Jaakkola)

   In the book, we discussed Bayesian inference of a Bernoulli rate parameter with the prior \( p(\theta) = \text{Beta}(\theta|\alpha, \beta) \). We know that, with this prior, the MAP estimate is given by
   \[
   \hat{\theta} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta + 2}
   \]  
   where \( N_1 \) is the number of heads, \( N_0 \) is the number of tails, and \( N = N_0 + N_1 \) is the total number of trials.

   1. Now consider the following prior, that believes the coin is fair, or is slightly biased towards tails:
      \[
      p(\theta) = \begin{cases} 
      0.5 & \text{if } \theta = 0.5 \\
      0.5 & \text{if } \theta = 0.4 \\
      0 & \text{otherwise} 
      \end{cases}
      \]  
      Derive the MAP estimate under this prior as a function of \( N_1 \) and \( N \).

   2. Suppose the true parameter is \( \theta = 0.41 \). Which prior leads to a better estimate when \( N \) is small? Which prior leads to a better estimate when \( N \) is large?

5. **Bayesian linear regression in 1d with known \( \sigma^2 \)**

   (Source: Bolstad)

   Consider fitting a model of the form
   \[
   p(y|x, \theta) = \mathcal{N}(y|w_0 + w_1x, \sigma^2)
   \]  
   to the data shown below:

   \[
   x = [94, 96, 94, 95, 104, 106, 108, 113, 115, 121, 131]; \quad y = [0.47, 0.75, 0.83, 0.98, 1.18, 1.29, 1.40, 1.60, 1.75, 1.90, 2.23];
   \]

   1. Compute an unbiased estimate of \( \sigma^2 \) using
      \[
      \hat{\sigma}^2 = \frac{1}{N-2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2
      \]  
      where \( \hat{y}_i = \hat{w}_0 + \hat{w}_1x_i \), where \( \hat{w} = (\hat{w}_0, \hat{w}_1) \) is the MLE.

   2. Now assume the following prior on \( w \):
      \[
      p(w) = p(w_0)p(w_1)
      \]  
      Use an (improper) uniform prior on \( w_0 \) and a \( \mathcal{N}(0, 1) \) prior on \( w_1 \). Show that this can be written as a Gaussian prior of the form \( p(w) = \mathcal{N}(w|w_0, V_0) \). What are \( w_0 \) and \( V_0 \)?

   3. Compute the marginal posterior of the slope, \( p(w_1|D, \sigma^2) \), where \( D \) is the data above, and \( \sigma^2 \) is the unbiased estimate computed above. What is \( E[w_1|D, \sigma^2] \) and \( \text{var}[w_1|D, \sigma^2] \)? Show your work. (You can use Matlab if you like.)

   4. What is a 95% credible interval for \( w_1 \)?