CSE 515T (Spring 2015) Assignment 2 solutions

1. (Curse of dimensionality.) Consider a d-dimensional, zero-mean, spherical multivariate Gaussian distribution:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I}_d).$$

Equivalently, each entry of $\mathbf{x}$ is drawn iid from a univariate standard normal distribution.

In familiar small dimensions ($d \leq 3$), "most" of the vectors drawn from a multivariate Gaussian distribution will lie near the mean. For example, the famous 68–95–99.7 rule for $d = 1$ indicates that large deviations from the mean are unusual. Here we will consider the behavior in larger dimensions.

- Draw 10 000 samples from $p(\mathbf{x})$ for each dimension in $d \in \{1, 5, 10, 50, 100\}$, and compute the length of each vector drawn: $y_d = \sqrt{\mathbf{x}^\top \mathbf{x}} = (\sum_i^d x_i^2)^{1/2}$. Estimate the distribution of each y_d using either a histogram or a kernel density estimate (in MATLAB, `hist` and `ksdensity`, respectively). Plot your estimates. (Please do not hand in your raw samples!) Summarize the behavior of this distribution as d increases.
- The true distribution of y_d^2 is a chi-square distribution with d degrees of freedom (the distribution of y_d itself is the less-commonly seen chi distribution). Use this fact to compute the probability that $y_d < 5$ for each of the dimensions in the last part.
- For $d = 1\,000$, compute the 5th and 95th percentiles of y_d. Is the mean $\mathbf{x} = \mathbf{0}$ a representative summary of the distribution in high dimensions? This behavior has been called "the curse of dimensionality."

Solution

Kernel density estimates of the empirical distributions of y (using 10 000 samples each) are shown in Figure 1. As the dimension increases, we see that the bulk of the probability mass actually lives in a thin "shell" centered around the origin: all samples are approximately the same length, with no vectors near the mean. This is somewhat unintuitive.

The compute the probability that $y_d < 5$, we can evaluate the corresponding χ^2 CDF at $y_d^2 = 25$:[1]

$$\begin{aligned}
\Pr(y < 5 \mid d = 1) &\approx 1.0000 \\
\Pr(y < 5 \mid d = 5) &\approx 0.9999 \\
\Pr(y < 5 \mid d = 10) &\approx 0.9947 \\
\Pr(y < 5 \mid d = 50) &\approx 0.0012 \\
\Pr(y < 5 \mid d = 100) &\approx 0.0000,
\end{aligned}$$

so the probability of being within a distance of five standard deviations form he mean decreases from near certainty to near impossibility, another surprising result.

For the last part, we invert the χ^2 CDF and take the square root.[2] The 5th percentile is 30.46 and the 95th percentile is 32.78. Again, we see that most of the mass lies in a narrow shell centered around the mean.

[1] Computed with `chi2cdf(25, [1, 5, 10, 50, 100])` in MATLAB.

[2] Computed with `sqrt(chi2inv([0.05, 0.95], 1000))` in MATLAB.

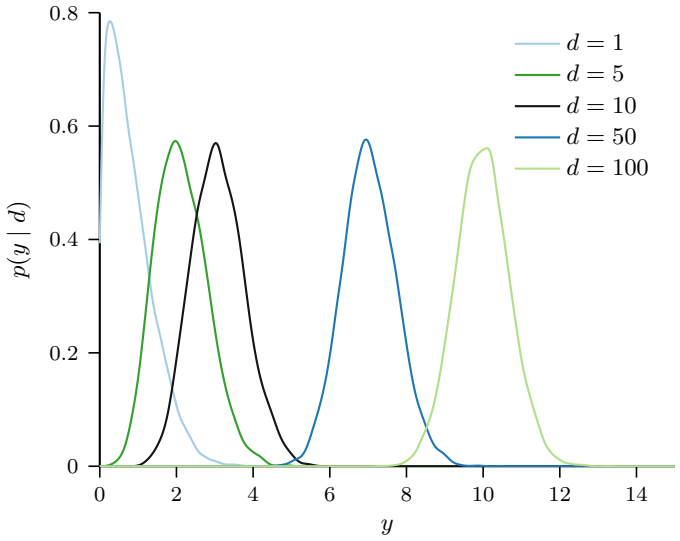

Figure 1: Empirical distributions of lengths y as a function of the dimension d.

Whether the mean is a representative summary is a much more complicated question with no definitive answer. In some sense, it's a very odd summary: in dimensions higher than 10 or so, we can't imagine seeing a vector anywhere near the mean! On the other hand, if we want to choose another single point to summarize the distribution, there's no clear better alternative. By definition, the mean minimizes the average squared distance from the chosen point to the vector $\mathbf{x}$. It just so happens that the *minimum* squared distance to the mean is relatively high in large dimension. This is the "curse of dimensionality:" all points are "far away from the mean" (and also each other!).

2. (Bayesian linear regression.) Consider the following data:

$$\mathbf{x} = [-2.26, -1.31, -0.43, 0.32, 0.34, 0.54, 0.86, 1.83, 2.77, 3.58]^\top;$$
$$\mathbf{y} = [1.03, 0.70, -0.68, -1.36, -1.74, -1.01, 0.24, 1.55, 1.68, 1.53]^\top.$$

Fix the noise variance at $\sigma^2 = 0.5^2$.

- Perform Bayesian linear regression for these data using the polynomial basis functions $\phi_k(x) = [1, x, x^2, \dots x^k]^\top$ for $k \in \{1, 2, 3\}$, in each case using the parameter prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$. Evaluate and plot the posterior means $\mathbb{E}[\mathbf{y}_* \mid \mathbf{X}_*, \mathcal{D}, \sigma^2]$ on the interval $x_* \in [-4, 4]$ for each model. Also plot the posterior mean plus-or-minus two times the posterior standard deviation:

  $$\mathbb{E}[\mathbf{y}_* \mid \mathbf{X}_*, \mathcal{D}, \sigma^2] \pm 2\sqrt{\mathrm{var}[\mathbf{y}_* \mid \mathbf{X}_*, \mathcal{D}, \sigma^2]}.$$

  This is a pointwise 95% credible interval for the regression function. Where is the pointwise uncertainty the largest?
- Compute the marginal likelihood of the data for each of the basis expansions above: $p(\mathbf{y} \mid \mathbf{X}, k, \sigma^2)$. Which model explains the data the best?

Solution

Given the feature expansions $\mathbf{\Phi} = \phi(\mathbf{X})$ and $\mathbf{\Phi}_* = \phi(\mathbf{X}_*)$, we may compute the posterior distribution of $\mathbf{y}_*$:

$$p(\mathbf{y}_* \mid \mathbf{X}_*, \mathcal{D}) = \mathcal{N}(\mathbf{y}_*; \mathbf{\Phi}_* \boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}}, \mathbf{X}_* \mathbf{\Sigma}_{\mathbf{w}|\mathcal{D}} \mathbf{X}_*^\top + \sigma^2 \mathbf{I}),$$

where

$$\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}} = \boldsymbol{\mu} + \mathbf{\Sigma}\mathbf{\Phi}^\top(\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{\Phi}\boldsymbol{\mu});$$
$$\mathbf{\Sigma}_{\mathbf{w}|\mathcal{D}} = \mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{\Phi}^\top(\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{\Phi}\mathbf{\Sigma}.$$

The diagonal of the posterior covariance for $\mathbf{y}*$, $\mathbf{\Phi}_*\mathbf{\Sigma}\mathbf{\Phi}_* + \sigma^2\mathbf{I}$, gives the desired variance for plotting the credible interval.

Plugging in the prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$ gives the result, plotted below. For all three models, the pointwise uncertainty is maximized on the extreme ranges of the domain at $x = -4$ and $x = 4$; we have few observations near either of these locations.

To compute the marginal likelihood, we must compute

$$p(\mathbf{y} \mid \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{y}; \mathbf{\Phi}\boldsymbol{\mu}, \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}).$$

Computing the marginal likelihood comes down to plugging in our observations $\mathbf{y}$ into this Gaussian PDF. In practice it is more convenient to compute the logarithm of the marginal likelihood, due to the large dynamic range of this function. For our data, we can compute:

$$\log p(\mathbf{y} \mid \mathbf{X}, \sigma^2, k = 1) = -32.9;$$
$$\log p(\mathbf{y} \mid \mathbf{X}, \sigma^2, k = 2) = -22.3;$$
$$\log p(\mathbf{y} \mid \mathbf{X}, \sigma^2, k = 3) = -22.2.$$

There is a clear preference for either the quadratic or the cubic model over the linear model, but there is no clear-cut winner between those two.

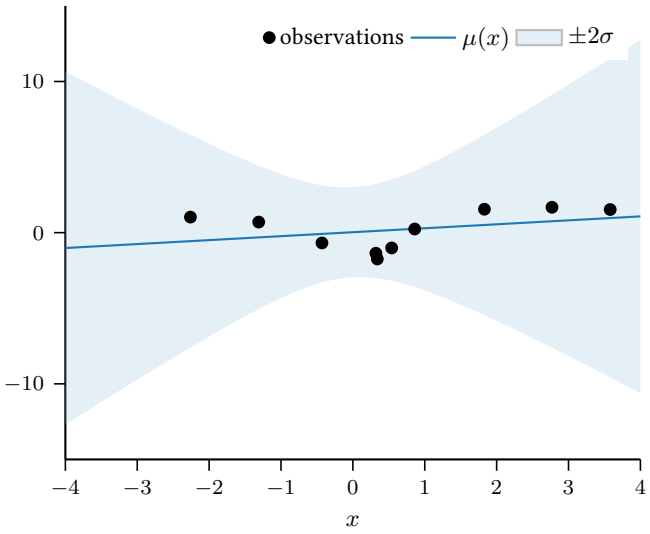

Figure 2: Posterior for $k = 1$.

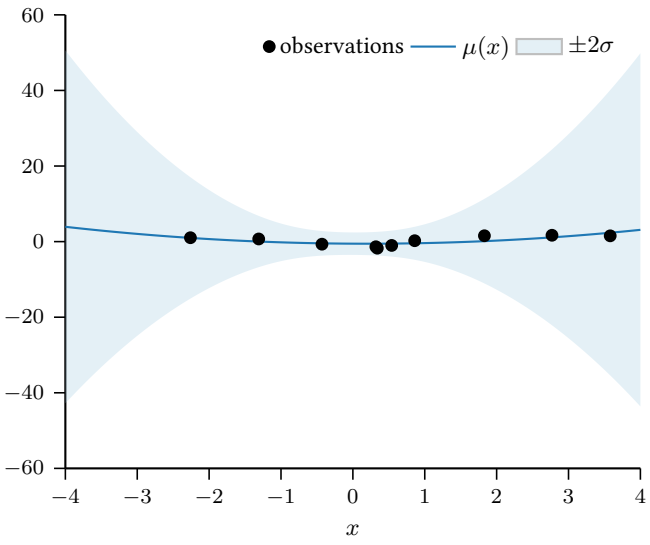

Figure 3: Posterior for $k = 2$.

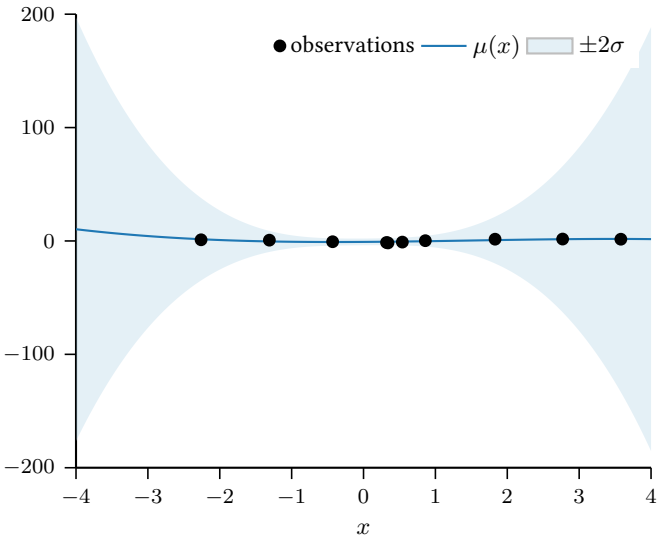

Figure 4: Posterior for $k = 3$.

3. (Optimal design for Bayesian linear regression.) Consider the data from the last problem, and suppose we have selected the quadratic model corresponding to $k = 2$ (do not assume that this is the answer to the last part of the last question). Imagine we are allowed to evaluate the function at a point x' of our choosing, giving a new dataset $\mathcal{D}' = \mathcal{D} \cup \{(x', y')\}$ and a new posterior for the parameters $p(\mathbf{w} \mid \mathcal{D}', \sigma^2) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}'}, \boldsymbol{\Sigma}_{\mathbf{w}|\mathcal{D}'})$. We hope to select the location x' to best improve our current model, under some quality measure.

   Assume that we ultimately wish to predict the function at a grid of points

$$\mathbf{x}_* = [-4, -3.5, -3, \ldots, 3.5, 4]^\top.$$

   We select the squared loss for a set of predictions $\hat{\mathbf{y}}_*$ at these points:

$$\ell(\mathbf{y}_*, \hat{\mathbf{y}}_*) = \sum_i \big((y_*)_i - (\hat{y}_*)_i\big)^2;$$

   therefore, we will predict using the new posterior mean $\hat{\mathbf{y}}_* = \mathbf{X}_* \boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}'}$.

   - Given a potential observation location x', derive a closed-form expression for the expected loss $\mathbb{E}\big[\ell(\mathbf{y}_*, \hat{\mathbf{y}}_*) \mid x', \mathcal{D}\big]$. Note: this does not require integration over y'! (What is the expected squared deviation from the mean?)
   - Plot the expected loss over the interval $x' \in [-4, 4]$. Where is the optimal location to sample the function?

   Note: this approach of actively selecting where to sample a function to maximize some utility function is known as *active learning* in machine learning and *optimal experimental design* in statistics. Bayesian decision theory provides a convenient and consistent framework for performing active learning with a variety of objectives.

Solution

Imagine for the sake of argument that we have been given a new observation (x', y') to our dataset $\mathcal{D}$, forming the augmented dataset $\mathcal{D}'$. Imagine further that we have computed the updated posterior $p(\mathbf{w} \mid \mathcal{D}', \sigma^2)$ with this new dataset.

We are compelled to predict the value of the function at the given test inputs $\mathbf{x}_*$. Under the given squared loss function $\ell(\mathbf{y}_*, \hat{\mathbf{y}}_*)$, we will predict the posterior mean $\hat{\mathbf{y}}_* \boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}'}$. (Recall the Bayes estimator under squared loss is the posterior mean).

Let us explicitly compute the expected loss given $\mathcal{D}$':

$$\begin{aligned}
\mathbb{E}\big[\ell(\mathbf{y}_*, \hat{\mathbf{y}}_*) \mid \mathbf{x}_*, \mathcal{D}'\big] &= \mathbb{E}\Big[\sum_i \big((y_*)_i - (\hat{y}_*)_i\big)^2 \mid \mathbf{x}_*, \mathcal{D}'\Big] \\
&= \sum_i \mathbb{E}\Big[\big((y_*)_i - \mathbb{E}\big[(y_*)_i \mid (x_*)_i, \mathcal{D}'\big]\big)^2 \mid (x_*)_i, \mathcal{D}'\Big] \\
&= \sum_i \operatorname{var}\big[(y_*)_i \mid (x_*)_i, \mathcal{D}'\big] \\
&= \operatorname{tr}\big(\boldsymbol{\Phi}_* \Sigma_{\mathbf{w}|\mathcal{D}'} \boldsymbol{\Phi}_*^\top + \sigma^2 \mathbf{I}\big),
\end{aligned}$$

where, in successive lines, we have: applied the linearity of expectation, substituted the posterior mean predictions for $\hat{\mathbf{y}}_*$, applied the definition of variance, and rewritten the sum of the variances in terms of the trace of the posterior covariance matrix over $\mathbf{y}_*$ given $\mathcal{D}'$.

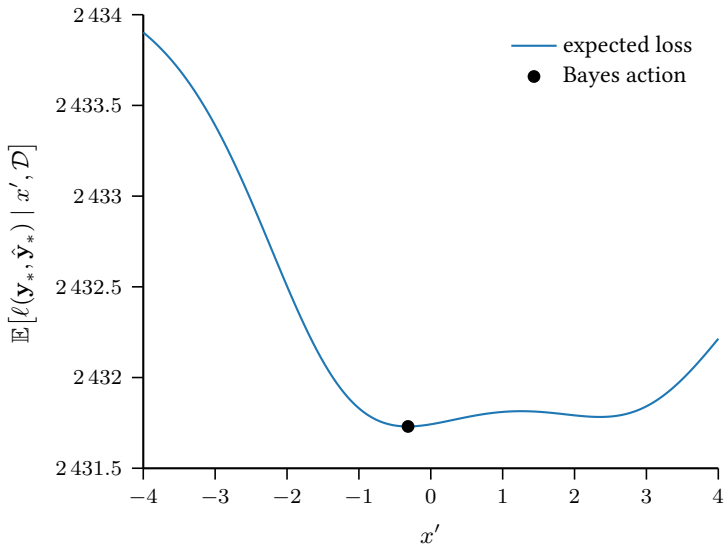

Figure 5: Expected loss for predicting $\mathbf{y}_*$ given $\mathcal{D}'$ as a function of x'.

The key observation here is that the posterior covariance matrix, and therefore the expected loss given $\mathcal{D}'$, *does not depend* on the value of y', only the location of the new input x'. So we may actually compute the future expected loss as a function of the next observation location x'. The Bayes action will then be to sample the function at the point minimizing the trace of the updated posterior covariance over $\mathbf{y}_*$.

In Figure 5, we plot the expected loss as a function of x'. The expected loss is minimized at $x' \approx -0.3163$, which is the Bayes action.

Of course, we could have iteratively performed this procedure to select every observation location! The result would fall under the general framework of so-called *active learning*.

4. (Woodbury matrix identity.) The *Woodbury matrix identity* is a very useful result. Let $\mathbf{A}$ be an $(n \times n)$ matrix, let $\mathbf{U}$ and $\mathbf{V}$ be $(n \times k)$ matrices, and let $\mathbf{C}$ be a $(k \times k)$ matrix. Then:

$$(\mathbf{A} + \mathbf{UCV}^\top)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{V}^\top\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^\top\mathbf{A}^{-1}.$$

This result is useful when you already have the inverse of a matrix $\mathbf{A}$ and want to know the inverse after a rank-k adjustment. When $k \ll n$, the Woodbry matrix identity can be considerably faster than direct inversion!

- Prove this result.
- Use this result to rewrite the posterior covariance of the weight vector $\mathbf{w}$ in Bayesian linear regression (as written in the notes to lecture 5) in a simpler form.

Solution

The first part of the problem can be completed by multiplying the right hand side by $(\mathbf{A} + \mathbf{UCV}^\top)$ and checking you get the identity.

The posterior covariance for $\mathbf{w}$ was previously given as

$$\mathbf{\Sigma}_{\mathbf{w}|\mathcal{D}} = \mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{X}^\top(\mathbf{X}\mathbf{\Sigma}\mathbf{X}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{\Sigma}.$$

Taking

$$\mathbf{A} = \mathbf{\Sigma}^{-1} \qquad \mathbf{U} = \mathbf{V} = \mathbf{X}^\top \qquad \mathbf{C} = \sigma^2\mathbf{I},$$

we may rewrite this as

$$\mathbf{\Sigma}_{\mathbf{w}|\mathcal{D}} = (\mathbf{\Sigma}^{-1} + \sigma^{-2}\mathbf{X}^\top\mathbf{X})^{-1}.$$

5. (Laplace approximation.) Find a Laplace approximation to the Gamma distribution:

$$p(\theta \mid \alpha, \beta) = \frac{1}{Z}\theta^{\alpha-1}\exp(-\beta\theta).$$

Plot the approximation against the true density for $(\alpha, \beta) = (2, 1/2)$.

The true value of the normalizing constant is

$$Z = \frac{\Gamma(\alpha)}{\beta^{\alpha}}.$$

If we fix $\beta = 1$, then $Z = \Gamma(\alpha)$, so we may use the Laplace approximation to estimate the Gamma function. Analyze the quality of this approximation as a function of α.

Solution

We first define the unnormalized log density $\Psi(\theta)$:

$$\Psi(\theta) = \log p(\theta \mid \alpha, \beta) = (\alpha - 1)\log\theta - \beta\theta.$$

Next we find the maximal value of the distribution, $\hat{\theta}$, by computing the derivative and setting to zero:

$$0 = \frac{d}{d\theta}\Psi(\theta) = \frac{\alpha - 1}{\theta} - \beta \quad \Rightarrow \quad \hat{\theta} = \frac{\alpha - 1}{\beta}.$$

Next we compute the negative Hessian of Ψ at $\hat{\theta}$. Note that here we have a one-dimensional density, so the Hessian is simply equal to the second derivative:

$$H = -\frac{d^2}{d\theta^2}\Psi(\theta)\bigg|_{\theta=\hat{\theta}} = \frac{\alpha - 1}{\theta^2}\bigg|_{\theta=\hat{\theta}} = \frac{\beta^2}{\alpha - 1}.$$

Now the Laplace approximation to the gamma distribution is

$$p(\theta \mid \alpha, \beta) \approx \mathcal{N}(\theta; \hat{\theta}, H^{-1}) = \mathcal{N}\left(\theta; \frac{\alpha - 1}{\beta}, \frac{\alpha - 1}{\beta^2}\right).$$

The corresponding estimate for the normalizing constant Z is

$$Z \approx \exp\big(\Psi(\hat{\boldsymbol{\theta}})\big)\sqrt{\frac{(2\pi)^d}{\det \mathbf{H}}} = p(\hat{\theta} \mid \alpha, \beta)\sqrt{\frac{2\pi}{H}} = \sqrt{\frac{2\pi(\alpha - 1)}{\beta^2}}\left(\frac{\alpha - 1}{\beta}\right)^{\alpha-1}\exp\big(-(\alpha - 1)\big).$$

Plugging in $\beta = 1$ and using the true normalizing constant of the gamma distribution, we have the approximation

$$\Gamma(\alpha) \approx \sqrt{2\pi}(\alpha - 1)^{\alpha - 1/2}\exp\big(-(\alpha - 1)\big).$$

Note we also have an approximation to the logarithm of Γ:

$$\log\Gamma(\alpha) \approx \tfrac{1}{2}\log 2\pi + (\alpha - 1/2)\log(\alpha - 1) - (\alpha - 1).$$

Figure 6 shows the resulting approximation to $Z = \Gamma(\alpha)$ as a function of α. The approximation appears to be quite good for $\alpha \geq 2$.

To those who are mathematically inclined, note that $n! = \Gamma(n + 1)$. With a bit of manipulation, we have actually rediscovered a very famous result known as *Stirling's approximation:*

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

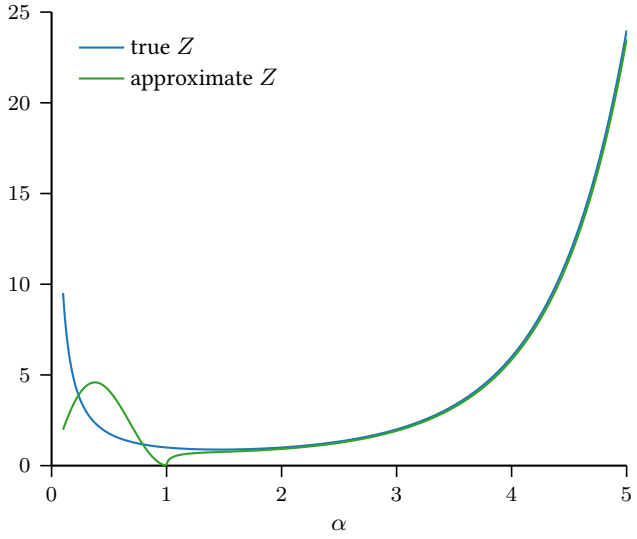

Figure 6: Laplace approximation to $Z = \Gamma(\alpha)$ as a function of α.