COMP 551 – Applied Machine Learning
Lecture 19: Bayesian Inference

Associate Instructor: Herke van Hoof (herke.vanhoof@mcgill.ca)

Class web page: *www.cs.mcgill.ca/~jpineau/comp551*

Slides

- Temporarily available at:

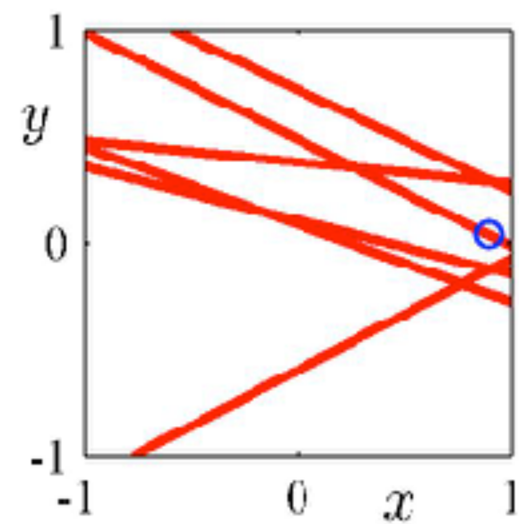
  http://cs.mcgill.ca/~hvanho2/media/19BayesianInference.pdf

- Quiz

  Will be online by tonight

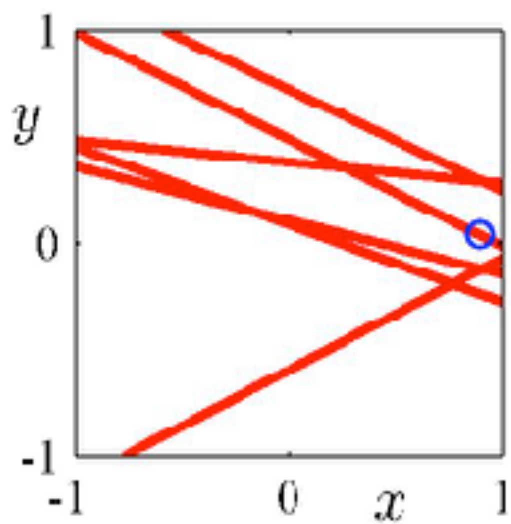
Bayesian probabilities

- An example from regression
- Given few noisy data points, multiple models conceivable
- Can we quantify uncertainty over models using probabilities?



Bayesian probabilities

- An example from regression
- Given few noisy data points, multiple models conceivable
- Can we quantify uncertainty over models using probabilities?
- Classical / frequentist statistics: no
  - Probability represents frequency of repeatable event
  - There is only one true model, we cannot observe multiple realisations of the true model



Copyright C.M. Bishop, PRML

Bayesian probabilities

- Bayesian view of probability
  - Uses probability to represent uncertainty
- Well-founded
  - When manipulating uncertainty, certain rules need to be respected to make rational choices
  - These rules are equivalent to the rules of probability

Goals of the lecture

At the end of the lecture, you are able to

- Formulate Bayesian view on probability
- Give reasons for (and against) Bayesian methods are used
- Understand Bayesian inference and prediction steps
- Give some examples with analytical solutions
- Use posterior and predictive distributions in decision making

Bayesian probabilities

- To specify uncertainty, need to specify a model
  - Prior over model parameters $p(\mathbf{w})$
  - Likelihood term $p(\mathcal{D}|\mathbf{w})$
- Dataset

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$$

- **Inference** using Bayes' theorem

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

Bayesian probabilities

- **Predictions**

$$p(y^*|\mathbf{x}^*, \mathcal{D}) = \int_{\mathbb{R}} p(y^*, \mathbf{w}|\mathbf{x}^*, \mathcal{D}) d\mathbf{w}$$

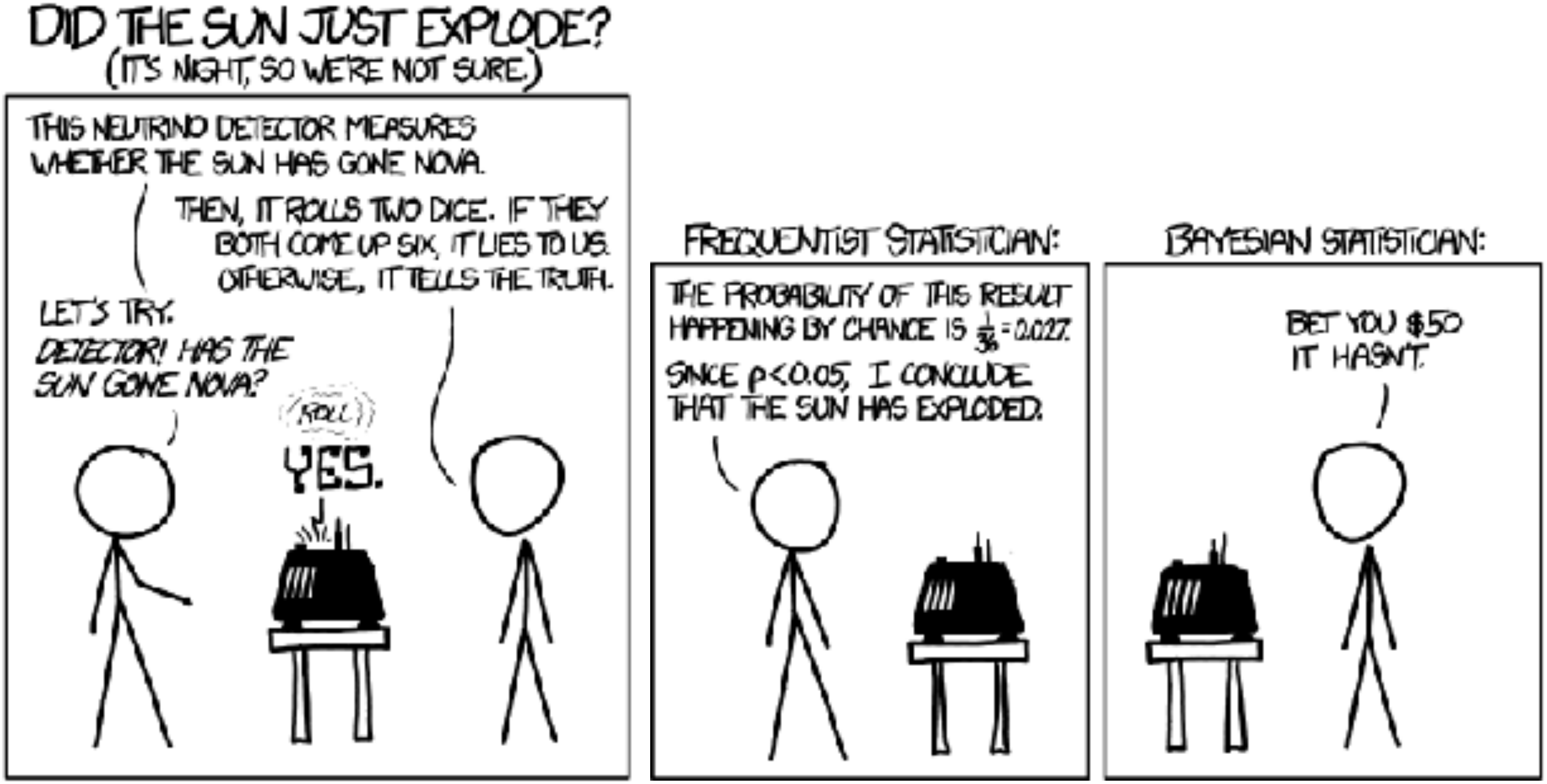
$$p(y^*|\mathbf{x}^*, \mathcal{D}) = \int_{\mathbb{R}^N} p(\mathbf{w}|\mathcal{D}) p(y^*|\mathbf{x}^*, \mathbf{w}) d\mathbf{w}$$

- Rather than fixing a fixed value for parameters, **integrate over all possible parameter values**!

Bayesian probabilities

- Note: **that Bayes' theorem is used does not mean a method uses a Bayesian view on probabilities!**
- Bayes' theorem is a consequence of the sum and product rules of probability
- Can relate the conditional probabilities of repeatable random events
  - Alarm vs. burglary
- Many frequentist methods refer to Bayes' theorem (naive Bayes, Bayesian networks)
- Bayesian view on probability: **Can represent uncertainty** (in parameters, unique events) **using probability**

Bayesian probabilities



Randall Munroe / xkcd.com

Why Bayesian probabilities?

- **Maximum likelihood estimates can have large variance**
  - Overfitting in e.g. linear regression models
  - MLE of coin flip probabilities with three sequential ‘heads’

Why Bayesian probabilities?

- Maximum likelihood estimates can have large variance
- **We might desire or need an estimate of uncertainty**
  - Use uncertainty in decision making

    Knowing uncertainty important for many loss functions
  - Use uncertainty to decide which data to acquire

    (active learning, experimental design)

Why Bayesian probabilities?

- Maximum likelihood estimates can have large variance
- We might desire or need an estimate of uncertainty
- **Have small dataset, unreliable data, or small batches of data**
  - Account for reliability of different pieces of evidence
  - Possible to update posterior incrementally with new data
  - Variance problem especially bad with small data sets

Why Bayesian probabilities?

- Maximum likelihood estimates can have large variance
- We might desire or need an estimate of uncertainty
- Have small dataset, unreliable data, or small batches of data
- **Use prior knowledge in a principled fashion**

Why Bayesian probabilities?

- Maximum likelihood estimates can have large variance
- We might desire or need an estimate of uncertainty
- Have small dataset, unreliable data, or small batches of data
- Use prior knowledge in a principled fashion
- **In practice, using prior knowledge and uncertainty particularly makes difference with small data sets**

Why not Bayesian probabilities?

- Prior induces bias
- Misspecified priors: if prior is wrong, posterior can be far off
- Prior often chosen for mathematical convenience, not actually knowledge of the problem
- In contrast to frequentist probability, uncertainty is subjective, different between different people / agents

Algorithms for Bayesian inference

- What do we need to do?
  - Dataset, e.g. $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$
  - Inference
    $$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$
  - Prediction
    $$p(y^*|\mathbf{x}^*, \mathcal{D}) = \int_{\mathbb{R}^N} p(\mathbf{w}|\mathcal{D})p(y^*|\mathbf{x}^*, \mathbf{w})d\mathbf{w}$$

- **When can we do these steps (in closed form)?**

Algorithms for Bayesian inference

- Inference

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

- Posterior can act like a prior

$$p(\mathbf{w}|\mathcal{D}_1, \mathcal{D}_2) = \frac{p(\mathcal{D}_2|\mathbf{w})p(\mathbf{w}|\mathcal{D}_1)}{p(\mathcal{D}_2)}$$

- Desirable that posterior and prior have same family!
  - Otherwise posterior would get more complex with each step
- Such priors are called conjugate priors to a likelihood function

Algorithms for Bayesian inference

- Prediction

$$p(y^*|\mathbf{x}^*, \mathcal{D}) = \int_{\mathbb{R}^N} p(\mathbf{w}|\mathcal{D})p(y^*|\mathbf{x}^*, \mathbf{w})d\mathbf{w}$$

same family as prior

- Argument of the integral is unnormalised distribution over **w**
- Integral calculates the normalisation constant
- For prior conjugate to likelihood function, constant is known

Algorithms for Bayesian inference

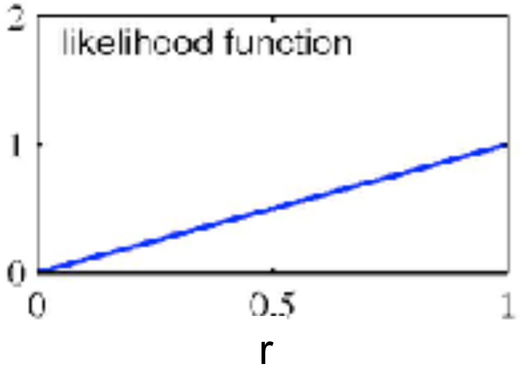
- Not all likelihood functions have conjugate priors
- However, so-called exponential family distributions do
  - Normal
  - Exponential
  - Beta
  - Bernoulli
  - Categorical
  - …

Simple example: coin toss

- Flip unfair coin
- Probability of 'heads' unknown value r
- Likelihood:

$$\mathrm{Bern}(x|r) = r^x(1-r)^{1-x}$$

  - x is one ('heads') or zero ('tails')
  - r is unknown parameter, between 0 and 1

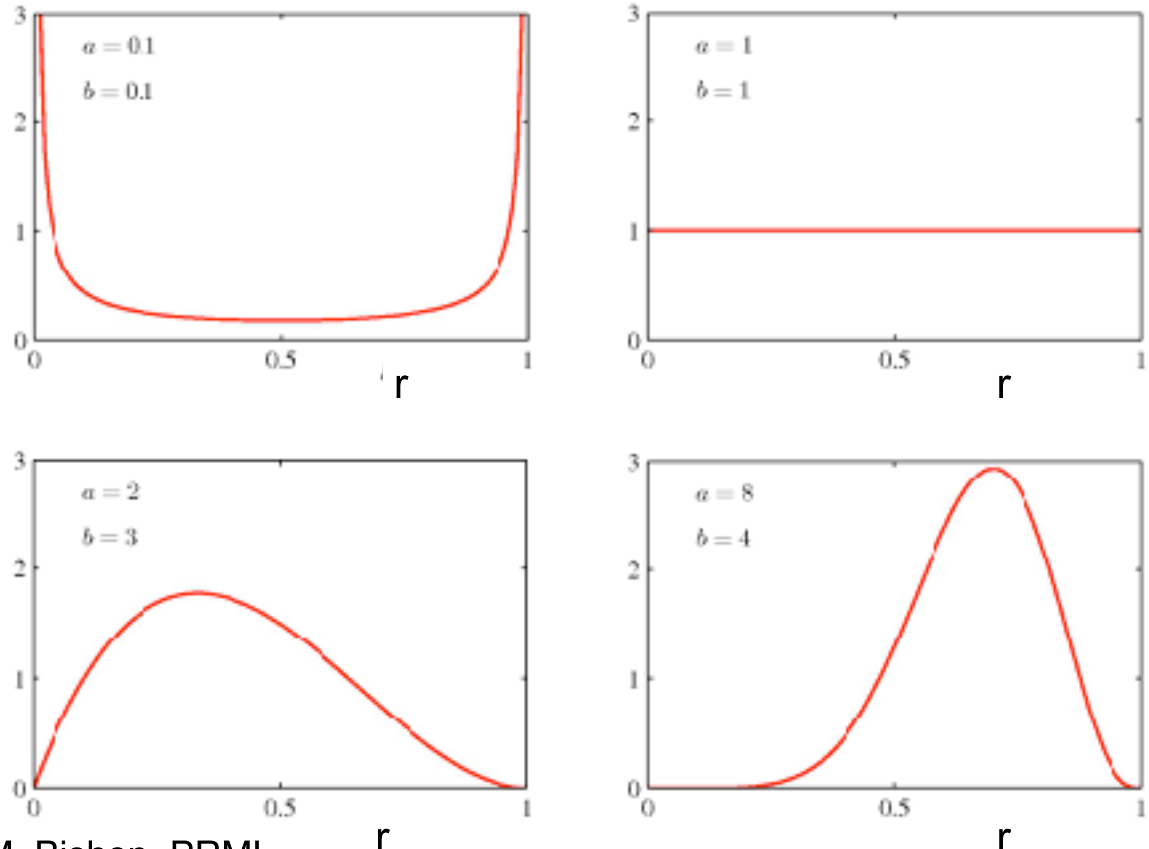


likelihood for x=1

Copyright C.M. Bishop, PRML

Simple example: coin toss

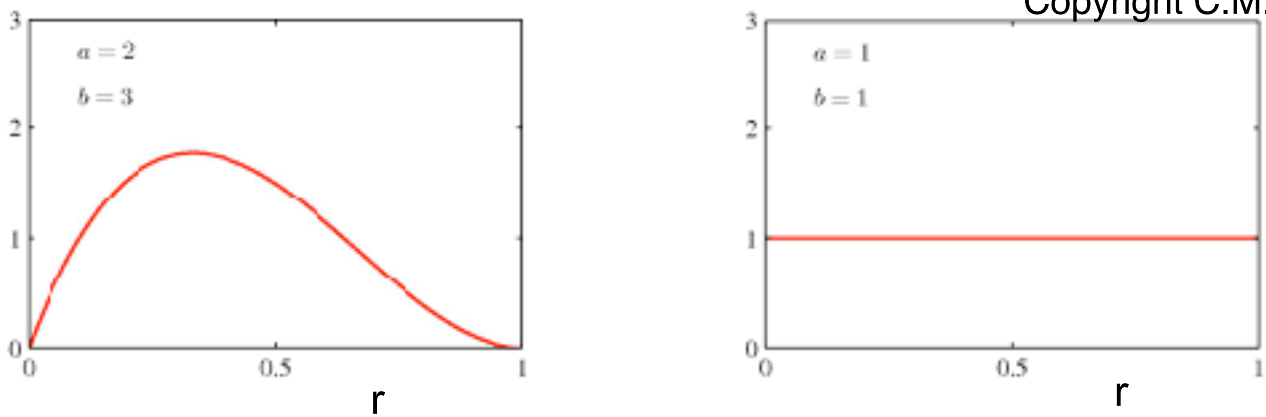
- Conjugate prior: $$\mathrm{Beta}(r|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} r^{a-1}(1-r)^{b-1}$$



Simple example: coin toss

- Conjugate prior: $$\mathrm{Beta}(r|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} r^{a-1}(1-r)^{b-1}$$

- Prior denotes a priori belief over the value r
- r is a value between 0 and 1 (denotes prob. of heads or tails)
- a, b are 'hyperparameters'



coin probably more likely to give 'tails' | no idea about the fairness

Simple example: coin toss

- Model:
  - Likelihood:

$$\mathrm{Bern}(x|r) = r^x(1-r)^{1-x}$$

  - Conjugate prior:

$$\mathrm{Beta}(r|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} r^{a-1}(1-r)^{b-1}$$

  - Posterior = prior x likelihood / normalisation factor
  - Note the similarity in the factors

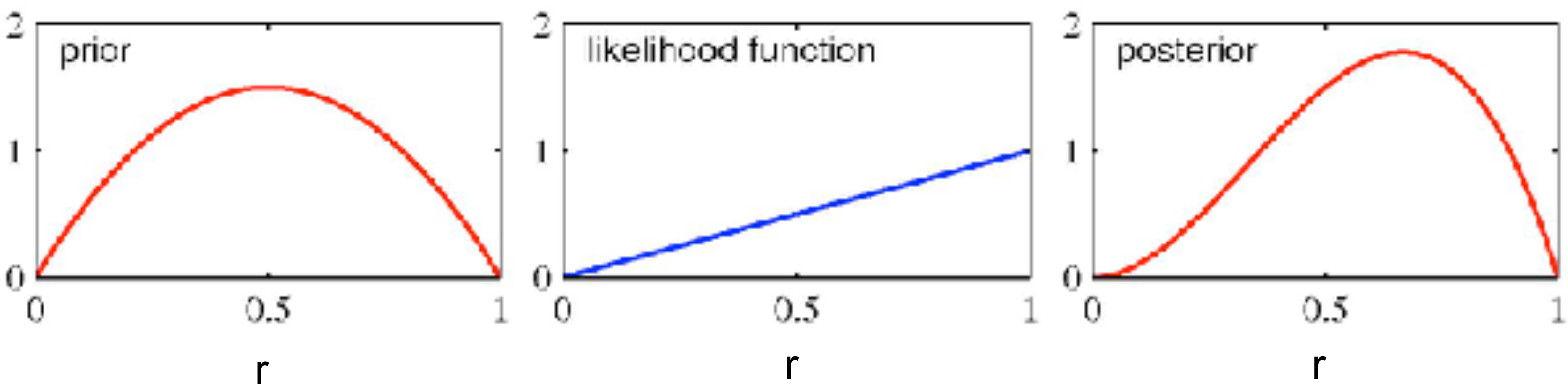
$$p(r|x) = z^{-1} r^{a+x-1}(1-r)^{b-x}$$

normalization factor

again beta distribution

Simple example: coin toss

- Posterior: $p(r|x) = z^{-1} r^{a+x-1} (1-r)^{b-x}$



- We observe more 'heads' -> suspect more strongly coin is biased
- Note that a, b get added to the actual outcome: 'pseudo-observations'
- Updated a,b can now be used as 'working prior' for the next coin flip

Simple example: coin toss

- Posterior: $p(r|x) = z^{-1} r^{a+x-1}(1-r)^{b-x}$
- Prediction: $p(x=1|\mathcal{D}) = \int_0^1 p(x=1|r)p(r|\mathcal{D})dr$

likelihood | posterior

$$= \frac{\#\text{heads} + a}{\#\text{heads} + \#\text{tails} + a + b}$$

- Instead of taking one parameter value, average over all of them
- a, b, again interpretable as effective # observations
- **Consider the difference if a=b=1, #heads=1, #tails=0**

Simple example: coin toss

- Posterior: $p(r|x) = z^{-1} r^{a+x-1} (1-r)^{b-x}$
- Prediction: $p(x=1|\mathcal{D}) = \int_0^1 p(x=1|r) p(r|\mathcal{D}) dr$

  likelihood | posterior

$$= \frac{\#\text{heads} + a}{\#\text{heads} + \#\text{tails} + a + b}$$

- Instead of taking one parameter value, average over all of them
- a, b, again interpretable as effective # observations
- Consider the difference if a=b=1, #heads=1, #tails=0
- Note that as #flips increases, prior starts to matter less

Simple example: coin toss

- Instead of taking one parameter value, average over all of them
  - True for all Bayesian models
- Hyperparameters interpretable as effective # observations
  - True for many Bayesian models
    (depends on parametrization)
- As amount of data increases, prior starts to matter less
  - True for all Bayesian models

Example 2: mean of a 1d Gaussian

- Try to learn the mean of a Gaussian distribution
- Model:
  - Likelihood $p(y) = \mathcal{N}(\mu, \sigma^2)$
  - Conjugate prior $p(\mu) = \mathcal{N}(0, \alpha^{-1})$
- Assume variances of the distributions are known
- We know the mean is close to zero but not its exact value

Example 2: inference for Gaussian

- From the shape of the distributions we see again some similarity:
  - log likelihood $$\text{const} - \frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}$$
  - log conjugate prior $$\text{const} - \frac{1}{2}\mu^2\alpha$$
- Now find log posterior

Inference for Gaussian

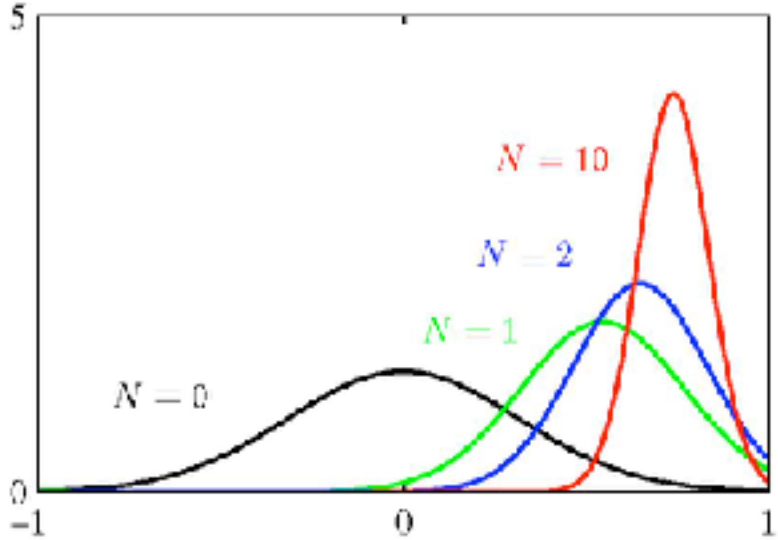
$$\text{const} - \frac{1}{2}\left(\frac{(y-\mu)^2}{\sigma^2} + \mu^2\alpha\right)$$

$$\begin{aligned}
\frac{(y-\mu)^2}{\sigma^2} + \mu^2\alpha &= -2\frac{y\mu}{\sigma^2} + \frac{\mu^2}{\sigma^2} + \mu^2\alpha + \text{const} \\
&= -2\frac{y\mu}{\sigma^2} + (\alpha + \sigma^{-2})\mu^2 + \text{const} \\
&= -2\frac{\alpha+\sigma^{-2}}{\alpha+\sigma^{-2}}\frac{1}{\sigma^2}y\mu + (\alpha + \sigma^{-2})\mu^2 + \text{const} \\
&= \frac{\left(\frac{\sigma^{-2}}{\alpha+\sigma^{-2}}y - \mu\right)^2}{(\alpha+\sigma^{-2})^{-1}} + \text{const}
\end{aligned}$$

mean of posterior distribution of μ: between MLE (y) and paprior (0)

covariance of posterior: smaller than either covariance of likelihood or prior

Inference for Gaussian



Prediction for Gaussian

- Prediction

$$p(y^*|\mathcal{D}) = \int_{-\infty}^{\infty} p(y^*, \mu|\mathcal{D}) d\mu$$

$$= \int_{-\infty}^{\infty} p(y^*|\mu) p(\mu|\mathcal{D}) d\mu$$

$$= \int_{-\infty}^{\infty} \mathcal{N}(y^*|\mu, \sigma^2) \mathcal{N}\left(\mu \,\middle|\, \frac{\sigma^{-2}}{\alpha + \sigma^{-2}} y_{\text{train}}, \frac{1}{\alpha + \sigma^{-2}}\right) d\mu$$

- Convolution of Gaussians, can be solved in closed form

$$p(y^*|\mathcal{D}) = \mathcal{N}\left(y^* \,\middle|\, \frac{\sigma^{-2}}{\alpha + \sigma^{-2}} y_{\text{train}}, \sigma^2 + \frac{1}{\alpha + \sigma^{-2}}\right)$$

noise + parameter uncertainty

Bayesian linear regression

- More complex example: Bayesian linear regression
- Model:
  - Likelihood
    $$p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2)$$
  - Conjugate prior
    $$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$$
  - Prior precision α and noise variance σ^2 considered known
  - Linear regression with uncertainty about the parameters