02477 – Bayesian Machine Learning: Lecture 2

Michael Riis Andersen

Technical University of Denmark,
DTU Compute, Department of Applied Math and Computer Science

Outline

1. Quick re-cap of last week
2. Probabilistic machine learning
3. The plug-in approximation
4. Grid approximations for non-conjugate models
5. Introduction to exercise: towards logistic regression

Quick re-cap of last week

Quick re-cap of Beta-binomial model and what's next

- *Bayes' rule* provides a systematic way to combine data with prior knowledge

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- *The beta-binomial model* is a *conjugate* model

$$p(\theta) = \mathrm{Beta}(\theta|a_0, b_0) \qquad (\textit{Prior})$$

$$p(y|\theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y} \qquad (\textit{Likelihood})$$

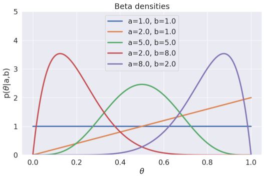
$$p(\theta|y) = \mathrm{Beta}(\theta|y+a_0, N-y+b_0) \qquad (\textit{Posterior})$$

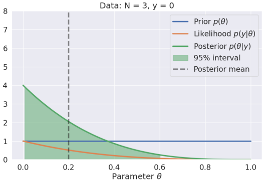
- Estimate θ using the *mean* of the *posterior distribution*

$$\theta_{\text{Bayes}} = \mathbb{E}\left[\theta|y\right] \equiv \int \theta\, p(\theta|y)\mathrm{d}\theta$$

- .. and use *credibility intervals* of the posterior to *quantify the uncertainty*

$$P(\theta \in [0.01, 0.60]\,|y) = 0.95$$



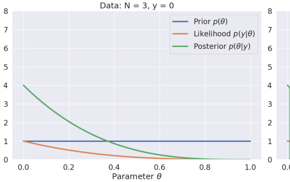


What about making predictions?

Example continued: suppose we have this website ad with $N = 3$ views and $y = 0$ clicks.

- Using a *uniform prior*, i.e. $p(\theta) = \text{Beta}(\theta|1,1)$

$$
\begin{aligned}
p(\theta) &= \text{Beta}(\theta|1,1) && (\textit{Prior}) \\
p(y|\theta) &= \binom{3}{0}\theta^0(1-\theta)^3 && (\textit{Likelihood}) \\
p(\theta|y) &= \text{Beta}(\theta|1,4) && (\textit{Posterior})
\end{aligned}
$$



- *Summarize* our knowledge using posterior

$$\mathbb{E}[\theta|y] = \frac{1}{5}, \qquad P(\theta \in [0.01, 0.60]\,|y) \approx 0.95$$

- **Goal**: *predict* number of clicks y^* in the next $N^* = 50$ views?

$$p(y^*|N^*,\theta) = \text{Bin}(y^*|N^*,\theta)$$

- Recall: mean of a Binomial random variable with prob. θ:

$$\mathbb{E}[y^*] = N^*\theta = 50\theta$$

Which value of θ to use? How do we use the *posterior knowledge* to make predictions? How do we take the *uncertainty* into account?

Probabilistic machine learning

Probabilistic machine learning I

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

- A probabilistic model is *completely specified* by its *joint distribution*
- Consider a model with two *random variables*: y (data) and θ (unknown parameter)
- The *joint distribution* of all random variables can be expressed via the *product rule*

$$p(\theta, y) = p(y|\theta)p(\theta)$$

- The *posterior distribution* can be obtained by *conditioning* on y

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- The *evidence* $p(y)$ can be obtained from the joint distribution via the *sum rule*

$$p(y) = \int p(y,\theta)\mathrm{d}\theta = \int p(y|\theta)p(\theta)\mathrm{d}\theta$$

- Hence, in theory, we can derive *all quantities of interest* from the *joint distribution*

Probabilistic machine learning II

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

- A probabilistic model is *completely specified* by its *joint distribution*

$$p(\theta, y) = p(y|\theta)p(\theta)$$

- What if we have more than one observed variable, e.g. y_1 and y_2?

- For a *broad class of models* the likelihood can be further decomposed using *conditional independence*:

$$p(y_1, y_2|\theta) = p(y_1|\theta)p(y_2|\theta)$$

- ... or more generally for $\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}$

$$p(\boldsymbol{y}|\theta) = p(y_1|\theta)p(y_2|\theta)\dots p(y_N|\theta) = \prod_{n=1}^{N} p(y_n|\theta)$$

- The *joint distribution* becomes

$$p(\theta, \boldsymbol{y}) = p(\boldsymbol{y}|\theta)p(\theta) = \prod_{n=1}^{N} p(y_n|\theta)p(\theta)$$

Website ad example continued: Making predictions

- **Example continued:** Your website ad has been shown $N = 123$ times and generated $y = 12$ clicks. Suppose you pay for another $N^* = 50$ views, how many clicks y^* should you expect *given the observed data*?

- Assuming each click can be modelled using *conditionally independent* Bernoulli trials with the *same probability* θ

$$p(y|\theta) = \text{Bin}(y|N, \theta)$$
$$p(y^*|\theta) = \text{Bin}(y|N^*, \theta)$$

- The assumption of *conditional independence* implies

$$p(y, y^*|\theta) = p(y|\theta)p(y^*|\theta) = \text{Bin}(y|N, \theta)\text{Bin}(y|N^*, \theta)$$

- Completing the model by *imposing a prior* for θ

$$p(\theta) = \text{Beta}(\theta|a_0, b_0)$$

- **Goal**: compute *predictive distribution* of y^* given we have observed $y = 12$, i.e. $p(y^*|y = 12)$.

A probabilistic perspective on making predictions

Product rule
$p(\boldsymbol{a}, \boldsymbol{b}) = p(\boldsymbol{b}|\boldsymbol{a})p(\boldsymbol{a})$

Sum rule
$p(\boldsymbol{b}) = \int p(\boldsymbol{a}, \boldsymbol{b})\mathrm{d}\boldsymbol{a}$

Conditional
$p(\boldsymbol{a}|\boldsymbol{b}) = \frac{p(\boldsymbol{a},\boldsymbol{b})}{p(\boldsymbol{b})}$

Conditional independence
$p(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{c}) = p(\boldsymbol{a}|\boldsymbol{c})p(\boldsymbol{b}|\boldsymbol{c})$

Goal: Given some data y, what can we say about a new observation y^*?

- Step 1: Formulate *joint distribution* for *all variables* of interests

$$p(y^*, y, \theta) = p(y^*, y|\theta)p(\theta) = p(y^*|\theta)p(y|\theta)p(\theta)$$

- Step 2: *Condition* on the *observed data* y

$$p(y^*, \theta|y) = \frac{p(y*, y, \theta)}{p(y)} = \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}$$

- Step 3: *Marginalize* out parameter θ using the *sum rule* to get the *posterior predictive distribution*

$$p(y^*|y) = \int p(y^*, \theta|y)\mathrm{d}\theta = \int \frac{p(y^*|\theta)p(y|\theta)p(\theta)}{p(y)}\mathrm{d}\theta = \int p(y^*|\theta)p(\theta|y)\mathrm{d}\theta = \mathbb{E}_{p(\theta|y)}\left[p(y^*|\theta)\right]$$

- **Key take-away**: To reason about *y^* given y*, we need to *average the likelihood* for y^* wrt. to the *posterior distribution* $p(\theta|y)$.

Quiz time

Take the quiz called

Lecture 2: Prior, likelihood, posterior, posterior predictive

to test your understanding.

Website example

- **Example continued:** Your website ad has been shown $N = 123$ times and generated $y = 12$ clicks. Suppose you pay for another $N^* = 50$ views, how many clicks y^* should you expect *given the observed data*?

- We already defined the model

$$\begin{aligned} p(y|\theta) &= \text{Bin}(y|N,\theta) && (\textit{Likelihood}) \\ p(y^*|\theta) &= \text{Bin}(y^*|N^*,\theta) && (\textit{Predictive likelihood}) \\ p(\theta) &= \text{Beta}(\theta|a_0,b_0) && (\textit{Prior}) \end{aligned}$$

- We know how to compute the *posterior distribution*

$$p(\theta|y) = \text{Beta}(\theta|y + a_0, N - y + b_0)$$

- Next, we want to compute the *posterior predictive distribution*

$$p(y^*|y) = \int p(y^*|\theta)p(\theta|y)\mathrm{d}\theta = \int \text{Bin}(y|N^*,\theta)\text{Beta}(\theta|y + a_0, N - y + b_0)\mathrm{d}\theta$$

- *Intuition*: Instead of plugging in a single value for the parameter estimate, we plug in all possible values for θ and weight the result according to $p(\theta|y)$

Website example

- Compute *posterior predictive distribution*

$$
\begin{aligned}
p(y^* = k|y) &= \int \mathrm{Bin}(y = k|N^*, \theta)\mathrm{Beta}(\theta|y + a_0, N - y + b_0)\mathrm{d}\theta \\
&= \int \binom{N^*}{k}\theta^k(1-\theta)^{N^*-k}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta \\
&= \binom{N^*}{k}\frac{1}{B(\alpha,\beta)}\int \theta^k(1-\theta)^{N^*-k}\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathrm{d}\theta && \text{(Linearity)} \\
&= \binom{N^*}{k}\frac{1}{B(\alpha,\beta)}\int \theta^{k+\alpha-1}(1-\theta)^{\beta+N^*-k-1}\mathrm{d}\theta && \text{(Simplify)} \\
&= \binom{N^*}{k}\frac{1}{B(\alpha,\beta)}\int \theta^{k+\alpha-1}(1-\theta)^{\beta+N^*-k-1}\mathrm{d}\theta
\end{aligned}
$$

- We recognize the terms in green as the *functional form* of a Beta density, and hence, we know how to compute the integral

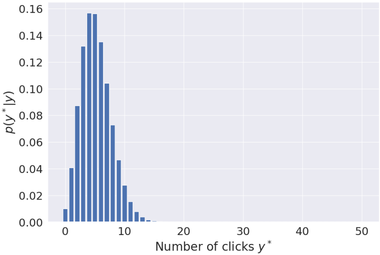
$$
p(y^* = k|y) = \binom{N^*}{k}\frac{B(\alpha+k,\beta+N^*-k)}{B(\alpha,\beta)}
$$

for $\alpha = y + a_0$ and $\beta = N - y + b_0$.

Website example: putting everything together

Example continued: Your website ad has been shown $N = 123$ times and generated $y = 12$ clicks. Suppose you pay for another $N^* = 50$ views, how many clicks y^* should you expect *given the observed data*?

$$\begin{aligned}
p(\theta) &= \mathrm{Beta}(\theta|1,1) && (\textit{Prior}) \\
p(y|\theta) &= \mathrm{Bin}(y|123,\theta) && (\textit{Likelihood}) \\
p(y^*|\theta) &= \mathrm{Bin}(y^*|50,\theta) && (\textit{Predictive likelihood}) \\
p(\theta|y) &= \mathrm{Beta}(\theta|13,112) && (\textit{Posterior})
\end{aligned}$$



- Distribution of clicks y^* based on views $N^* = 50$ views

$$p(y^* = k|y) = \binom{N^*}{k} \frac{B(\alpha + k, \beta + N^* - k)}{B(\alpha, \beta)} = \binom{50}{k} \frac{B(13 + k, 162 - k)}{B(13, 112)}$$

- The expected number of clicks given the data is

$$\mathbb{E}_{p(y^*|y)}[y^*] = \sum_{k=0}^{50} k p(y^* = k|y) \approx 5.2$$

The plug-in approximation

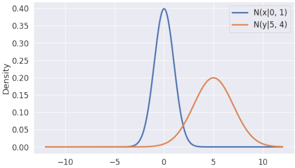
A few prerequisites first

Univariate Gaussians

- The *normal distribution* (also known as the Gaussian) is distribution over $x \in \mathbb{R}$ with density

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Two parameters: $\mu \equiv \mathbb{E}[x]$ and $\sigma^2 \equiv \mathbb{V}[x]$



- Widely popular due to Central limit theorems, maximum entropy principle, relation to least squares minimization, nice mathematical properties

- We will talk more about Gaussians later in this course

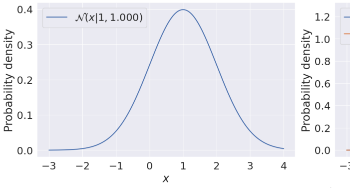
A few prerequisites first

Dirac's delta function

- Consider a Gaussian random variable $x \sim \mathcal{N}(\mu, \sigma^2)$.
  In the *limit* $\sigma^2 \to 0$, x is effectively a *constant* $x = \mu$:

- We say that x follows a *Dirac's delta distribution* centered at μ

$$p(x) = \lim_{\sigma^2 \to 0} \mathcal{N}(x|\mu, \sigma^2) = \delta(x - \mu)$$



- Important properties

$$\delta(x - \mu) = \begin{cases} \infty & \text{if } x = \mu \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\int \delta(x - \mu) \mathrm{d}x = 1 \tag{2}$$

$$\int f(x)\delta(x - \mu) \mathrm{d}x = f(\mu) \tag{3}$$

- Eq. (3) is called the *sifting property* and implies

$$\mathbb{E}_{\delta(x-\mu)}\left[f(x)\right] = f(\mu)$$

The plugin approximation

- We showed that the rules of probability theory dictates that

$$p(y^*|y) = \int p(y^*|\theta)p(\theta|y)\mathrm{d}\theta$$

- While this is *optimal* given the model, it can be *non-trivial* in practice

- If we *assume* that there is a *single best parameter* $\hat{\theta}$, e.g. $\hat{\theta}_{\text{MLE}}$ or $\hat{\theta}_{\text{MAP}}$, then we can approximate $p(\theta|y)$ using a *Dirac's delta function* $\delta(\theta - \hat{\theta})$

$$p(\theta|y) \approx \delta(\theta - \hat{\theta})$$

- Using the *sifting property* of Dirac's delta implies

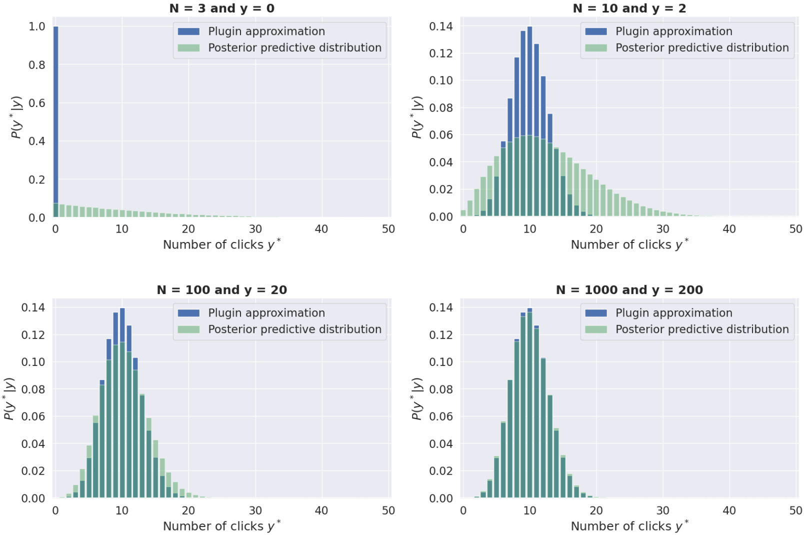
$$p(y^*|y) = \int p(y^*|\theta)p(\theta|y)\mathrm{d}\theta \approx \int p(y^*|\theta)\delta(\theta - \hat{\theta})\mathrm{d}\theta = p(y^*|\hat{\theta})$$

- Therefore, this is called a *plug-in approximation*.

- Very *easy to compute*, but *ignores uncertainty* for our estimate of θ, and hence, often *producing overconfident predictions*.

- This is how we make predictions in deep learning...

The posterior predictive distribution and plugin approximations



Grid approximations for non-conjugate models

Introduction to non-conjugate models

Big picture so far...

- We studied the binomial model for estimating proportions and imposed a Beta prior for θ for Bayesian inference
- We derived the *posterior* and *posterior predictive* distributions *analytically*. This is possible due to *conjugacy* of the Beta prior and binomial likelihood
- Supposed our analysis required computing the posterior mean for a *different prior*, e.g. $p(\theta) = \frac{1}{Z} e^{\sin(\pi\theta^2)}$

$$\mathbb{E}[\theta|y] = \int \theta p(\theta|y) \mathrm{d}\theta = \int \theta \frac{p(y|\theta)p(\theta)}{p(y)} \mathrm{d}\theta$$

- To compute the posterior mean, variance etc. we need the evidence $p(y)$

$$\begin{aligned} p(y) &= \int p(\theta|y)p(\theta)\mathrm{d}\theta = \int \mathrm{Bin}(y|N,\theta)\frac{1}{Z}e^{\sin(\pi\theta^2)}\mathrm{d}\theta \\ &= \int \binom{N}{y}\theta^y(1-\theta)^{N-y}\frac{1}{Z}e^{\sin(\pi\theta^2)}\mathrm{d}\theta \\ &= \binom{N}{y}\frac{1}{Z}\int \theta^y(1-\theta)^{N-y}e^{\sin(\pi\theta^2)}\mathrm{d}\theta = ? \end{aligned}$$

- Unfortunately, we *cannot* evaluate the evidence, i.e. $p(y)$ analytically *intractable* for most models of practical interest...

Approximate inference methods

- **In this course**: We will study several computational tools and approximation inference methods for dealing with such *intractable distributions*

  1. Grid approximations
  2. Laplace approximations
  3. Variational inference
  4. Markov Chain Monte Carlo methods

- **Goal for these methods**: Compute *tractable approximation* $q(\theta)$ of true posterior $p(\theta|y)$ such that

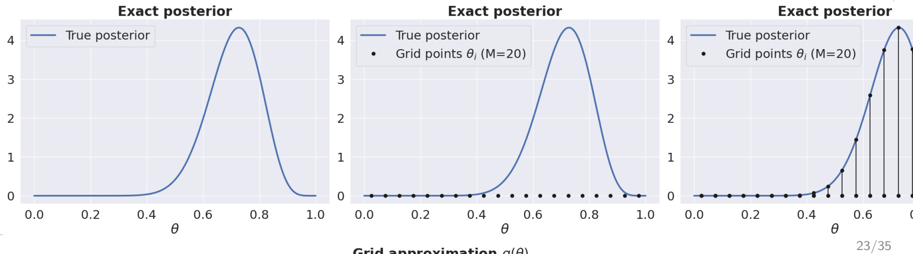
  1. $q(\theta)$ resembles the true posterior, i.e. $p(\theta|y) \approx q(\theta)$
  2. $q(\theta)$ should be tractable s.t. we can compute posterior summaries, predictions etc.

- **This week**: We will focus on *grid approximations*, which are easy to understand and apply, and they will help build our intuition about marginalization.

The grid approximation

- **Constructing the grid approximation for $p(\theta|y)$**

  1. We define a set of grid points for θ: $0 \leq \theta_1 < \theta_2 < \cdots < \theta_M \leq 1$
  2. We evaluate the exact posterior (up to a constant) at all the grid points, i.e.
  $$\tilde{\pi}_i \propto p(\theta_i|y) \propto p(y|\theta_i)p(\theta_i)$$
  3. Sum all values to get normalization constant $Z = \sum_{i=1}^{M} \tilde{\pi}_i$
  4. Compute normalized probabilities $\pi_i = \frac{1}{Z}\tilde{\pi}_i$ to get the grid approximation
  $$q(\theta) = \sum_{i=1}^{M} \pi_i \delta(\theta - \theta_i)$$



The grid approximation

Posterior summaries

$$q(\theta) = \sum_{i=1}^{M} \pi_i \delta(\theta - \theta_i)$$

- The grid approximation is a discrete distribution, so computing summaries is easy, e.g the posterior mean

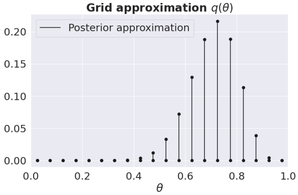
$$\mathbb{E}_{p(\theta|y)}\left[\theta\right] \approx \mathbb{E}_{q(\theta)}\left[\theta\right] = \sum_{i=1}^{M} \theta_i \pi_i$$

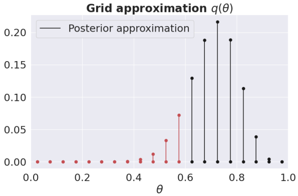
- ... and general expectations of $f(\theta)$

$$\mathbb{E}_{p(\theta|y)}\left[f(\theta)\right] \approx \mathbb{E}_{q(\theta)}\left[f(\theta)\right] = \sum_{i=1}^{M} f(\theta_i) \pi_i$$

- Example: Computing post. probabilities for $\theta < 0.6$

$$\begin{aligned} p(\theta < 0.6|y) &\approx q(\theta < 0.6) \\ &= \sum_{i=1}^{M} \mathbb{I}\left[\theta_j < 0.6\right] \pi_j \\ &= \sum_{i=1}^{j} \pi_i, \quad j = \max\left\{i|\theta_i < 0.6\right\} \end{aligned}$$





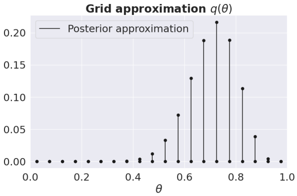
The grid approximation

The posterior predictive distribution

- General expectations of $f(\theta)$

$$\mathbb{E}_{p(\theta|y)}\left[f(\theta)\right] \approx \mathbb{E}_{q(\theta)}\left[f(\theta)\right] = \sum_{i=1}^{M} f(\theta_i)\pi_i$$

$$q(\theta) = \sum_{i=1}^{M} \pi_i \delta(\theta - \theta_i)$$



- The posterior predictive distribution $p(y^*|y)$

$$p(y^* = k|y) = \mathbb{E}_{p(\theta|y)}\left[\text{Bin}(y^* = k|N^*, \theta)\right]$$

- Hence, setting $f(\theta) = \text{Bin}(y^* = k|N^*, \theta_i)$ yields

$$p(y^* = k|y) \approx \sum_{i=1}^{M} \text{Bin}(y^* = k|N^*, \theta_i)\pi_i$$

- To make predictions we literally compute a weighted sum of all possible parameter values

The grid approximation

A few practical considerations

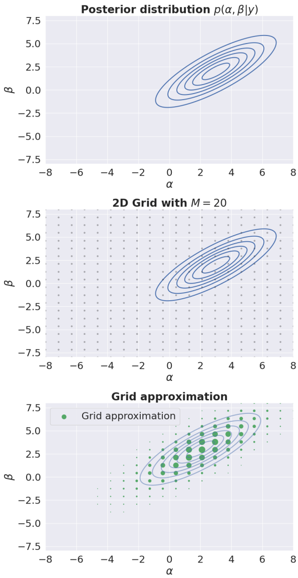
- Choosing the grid range
    - Range for $\theta \in [0, 1]$ is easy
    - For a different parameter $\alpha \in \mathbb{R}$, we need to choose an interval $[a, b]$ for the grid. Often identified visually.
- Scaling with model dimensionality
    - Suppose we use $M = 20$ points for each dimension

      1D: 20 evaluations

      2D: $20^2 = 400$ evaluations

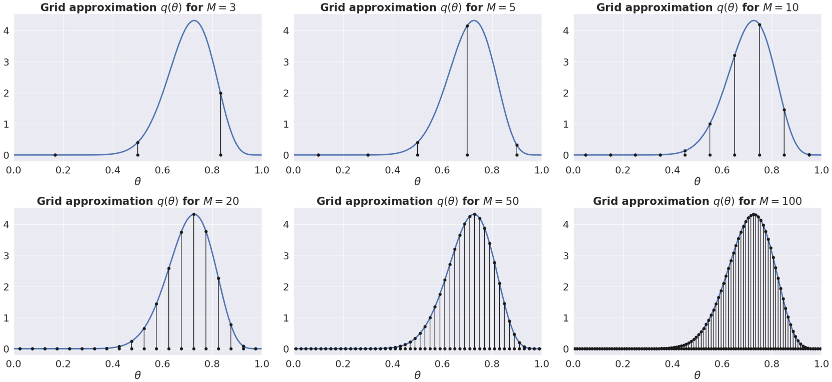
      3D: $20^3 = 8000$ evaluations

      4D: $16000^4 = 160000$ evaluations
    - Grid approximations do not scale well beyond 3-4 dimensions
- Number of grid points M
    - M is balance between computational cost and accuracy
    - Grid approximation is zero when evaluated outside the grid points
    - Often diminishing returns as M increases (next slide)



The grid approximation

Summary



- **Pros:** Simple, easy and intuitive.
- **Cons:** Suffers from curse of dimensionality and does not scale beyond 3-4 dimensions

Introduction to exercise: towards logistic regression

Towards logistic regression

- So far we focussed on modelling *proportions*, i.e. $\theta \in [0, 1]$ given data about y successes in N *conditionally independent trials*

- The binomial likelihood is also often used in *dose-response* models, which is key for determining "safe" dosages for drugs, pollution, foods etc.

- **Example**: A company wants to study side effects of their new drug

x (**Dose in mg**)	y (**# side effects**)	N (**# patients**)
80	0	69
160	4	832
320	13	835
480	20	459
640	12	324
800	6	103

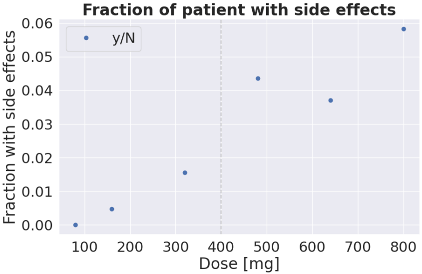
- We could analyze the data for each dose independently using a beta-binomial model, but we would like to ...

  1. understand how dose affect the probability of side effects
  2. make predictions for new dosages x^*
  3. borrow "statistical strength" across dosages

Towards logistic regression

Example & motivation II

x (**Dose in mg**)	y (**# side effects**)	N (**# patients**)	**y/N**
80	0	69	0
160	4	832	0.005
320	13	835	0.016
480	20	459	0.044
640	12	324	0.037
800	6	103	0.058



- How accurate can we predict the probability of side effects for $x^* = 400mg$?

Towards logistic regression
Setting up the likelihood

- For each dose x_i, we assume

$$y_i|x_i \sim \text{Bin}(y_i|N_i, \theta_i), \quad \theta_i \in [0, 1]$$

- We model the probability θ_i as function of the dose x_i, i.e.

$$\theta_i \equiv \theta(x_i) = \sigma(\alpha + \beta x),$$

  where $\sigma(x) : \mathbb{R} \rightarrow [0, 1]$ is a sigmoid function and $\alpha, \beta \in \mathbb{R}$ are model parameters.

- The likelihood of the i'th observation (x_i, y_i)

$$p(y_i|x_i, \alpha, \beta) = \text{Bin}(y_i|N_i, \theta_i)$$

- Assuming conditional independence we can write the joint likelihood

$$p(\mathbf{y}|\mathbf{x}, \alpha, \beta) = \prod_{i=1}^{M} p(y_i|x_i, \alpha, \beta) = \prod_{i=1}^{M} \text{Bin}(y_i|N_i, \theta_i),$$

  where $\mathbf{y} = [y_1, y_2, \ldots y_6]$ and similar for $\mathbf{x} = [x_1, x_2, \ldots x_6]$

- The predictive likelihood for y^* is

$$p(y^*|x^*, \alpha, \beta) = \text{Bin}(y^*|N_i, \theta^*)$$

  where $\theta^* \equiv \theta(x^*)$

Towards logistic regression

Setting up the prior

- The model parameters are α (intercept) and β (slope) of the generalized linear model

- Prior information: we have no prior information about the sign of the parameters. Hence, we choose a zero-mean Gaussian distributions

$$p(\alpha, \beta) = \mathcal{N}(\alpha|0, \sigma_\alpha^2)\mathcal{N}(\beta|0, \sigma_\beta^2), \qquad \sigma_\alpha^2, \sigma_\beta^2 > 0$$

- We can now write the *joint distribution* of $\alpha, \beta, \mathbf{y}, y^*$ using the *product rule*

$$p(\mathbf{y}, y^*, \alpha, \beta|\mathbf{x}, x^*) = p(y^*|x^*, \alpha, \beta)p(\mathbf{y}|\mathbf{x}, \alpha, \beta)p(\alpha, \beta)$$

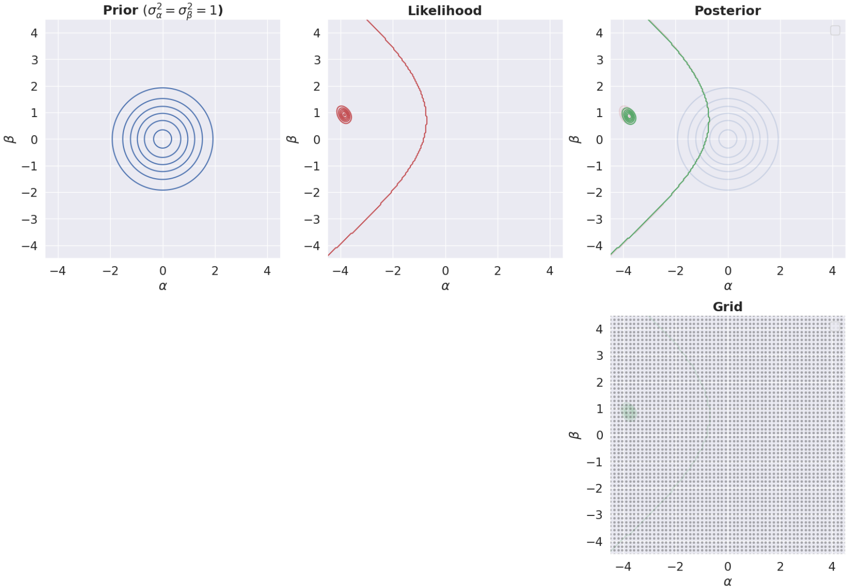
- The *posterior predictive distribution* is by *conditioning on* $\mathbf{y}$ and *marginalizing out the parameters* via the sum rule

$$p(y^*|\mathbf{y}, \mathbf{x}, x^*) = \iint p(y^*, \alpha, \beta|\mathbf{y}, \mathbf{x}, x^*)\mathrm{d}\alpha\mathrm{d}\beta = \iint \underbrace{p(y^*|x^*, \alpha, \beta)}_{\text{likelihood for } y^*} \underbrace{p(\alpha, \beta|\mathbf{y}, \mathbf{x})}_{\text{posterior distribution}} \mathrm{d}\alpha\mathrm{d}\beta$$

- After obtaining the posterior of α, β, we can *propagate* the posterior uncertainty of the parameters to any quantity that depends on α, β, i.e. $\theta(x) = \sigma(\alpha + \beta x)$, the fraction of people with side effects y^*/N etc.

Towards logistic regression

Visualizing the distributions



Towards logistic regression

Making predictions

Computing the posterior predictive distribution $p(y^* = k|\mathbf{y}, \mathbf{x}, x^*)$ **for** $x^* = 400\text{mg}$ **and** $N^* = 500$

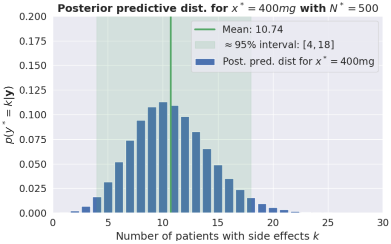
$$p(y^* = k|\mathbf{y}, \mathbf{x}, x^*) = \iint \underbrace{p(y^* = k|x^*, \alpha, \beta)}_{\text{likelihood for } y^*} \underbrace{p(\alpha, \beta|\mathbf{y}, \mathbf{x})}_{\text{posterior distribution}} \, \mathrm{d}\alpha \mathrm{d}\beta \qquad \text{(Sum rule)}$$

$$= \mathbb{E}_{p(\alpha,\beta|\mathbf{y},\mathbf{x},x^*)}\left[p(y^* = k|x^*, \alpha, \beta)\right] \qquad \text{(Integrals as expectation)}$$

$$= \mathbb{E}_{p(\alpha,\beta|\mathbf{y},\mathbf{x},x^*)}\left[\text{Bin}(y^*|N^*, \theta^*)\right] \qquad \text{(Inserting dist.)}$$

$$\approx \mathbb{E}_{q(\alpha,\beta)}\left[\text{Bin}(y^*|N^*, \theta^*)\right] \qquad \text{(Grid approx.)}$$

$$= \sum_{i,j} \text{Bin}(y^*|N^*, \sigma(\alpha_i + \beta_j x^*))\pi_{ij}$$



Intro to exercise

- On DTU Learn you will find an exercise for each week in notebook format
- We will spend all 4 hours from 13-17 working with the exercises
- In this exercise you will
  - Dive deeper into the Bayesian framework
  - Study and implement the probabilistic model for logistic regression for the Challenger Distaster dataset
  - Study and implement the grid approximations
  - Practice probabilistic reasoning
- Mix of pen&paper, programming and discussion questions
- Feel free to collaborate with your peers
- Ask for help!
  - Ask for help when stuck
  - Use teachers/TAs to check your understanding
  - Engage in discussion to practice
- Feedbacks persons: Meet at 16:45