Bayesian Inference

Viktor

October 15, 2020

1 Graph theory

We recap the basics of graph theory. [14, **?**]

Graph. A graph $G = (V, E)$ consistes of a set $V = \{1, \dots, m\}$ of vertices and a set $E \subseteq V \times V$ of edges. So each edge consists of a pair of vertices $(s, t) \in E$. For an *undirected* graph there is no distinction between (s, t) and (t, s). In a *directed* graph the edge orientation is distinguished, which is emphasised by the notation $(s \to t) := (s, t)$.

Remark: By definition a graph does not contain self-loops as edges $(s, s) \notin E,\ \forall s \in V$ nor does it contain multiple copies of the same vertex. This may be included within the framework of multigraphs.

Subgraph. A subgraph $G' = (V', E')$ of a given graph $G = (V, E)$ is a graph such that $V' \subseteq V$ and $E' \subseteq E$. Given a vertex subset $V' \subseteq V$ of $G = (V, E)$, the *vertex-induced subgraph* $G'(V') = (V', E(V'))$ has the edge set $E(V') = \{(s, t) \in E | s, t \in V'\}$. Given an edge subset $E' \subseteq E$ of $G = (V, E)$, the *edge-induced subgraph* $G'(E') = (V(E'), E')$ has vertex set $V(E') = \{s \in V | (s, t) \in E'\}$

Path. A path $P = (V(P), E(P))$ is a graph with vertex set $V(P) = \{v_0, \dots v_k\}$ and edge set $V(P) = \{(v_0, v_1), (v_1, v_2) \dots (v_{k-1}, v_k)\}$. The path joins vertex v_0 to vertex v_k. Of special interest are paths that are subgraphs of a given Graph $G = (V, E) :\ V(P) \subseteq V,\ E(P) \subseteq E$.

Cycle. A cylcle $C = (V(C), E(C))$ is a graph with vertex set $V(C) = \{v_0, \dots v_k\}$ and edge set $E(C) = \{(v_0, v_1), (v_1, v_2) \dots (v_{k-1}, v_k), \dots (v_k, v_0)\}$. An undirected graph is acyclic if it contains no cycles.

Bipartite. A graph $G = (V, E)$ is bipartite if its vertex set can be partitioned as a disjoint union $V = V_a \dot{\cup} V_b$ such that $(s, t) \in E \Rightarrow s \in V_a, t \in V_b$ (or vice versa).

Clique. A clique of a graph $G = (V, E)$ is a subset of vertices $V' \subseteq V$ that are all joined by vertices, $(s, t) \in E \ \forall s, t \in V'$. A clique V' is maximal if there is no vertex $v \in V \setminus V'$ such that $V \cup \{v\}$ is a clique.

Remark: Sometimes maximal cliques are just called cliques and non-maximal cliques are calle cliquos.

Chord. Given a cylce C with vertex set $V(C) = \{v_0, \ldots v_k\}$ and edge set $E(C) = \{(v_0, v_1), (v_1, v_2) \ldots (v_{k-1}, v_k), \ldots (v_k, v_0)\}$. A chord is an edge that is not part of $E(C)$. Given a Graph $G = (V, E)$ and a cycle C of length four or greater. C is *chordless* if the edge set E of G contains no chords for C.

Triangulated. A graph is triangulated if it contains no chordless cycles (of length four or greater).

Connected component. A connected component of a graph $G = (V, E)$ is a subset of vertices $V' \subseteq V$ such that $\forall \ s, t \in V'$, there exists a path in G joining s to t. A graph is *singlyconnected* if it consists of a single connected component.

Tree. A tree is an acyclic singly connected graph. It can be shown that a tree with m verices must have $m - 1$ edges.

Forest. A forest is an acyclic graph consisting of one or more connected components.

Hypergraph. A hypergraph is $G = (V, E)$ consists of a vertex set $V = \{1, 2, ..., m\}$, and a set E of hyperedges, with $E \ni h \subset V$. So, each hyperedge is a particular subset of V. In particular, an ordinary graph is a hypergraph with $|h| = 2$.

Factor graph. Given a hypergraph $G = (V, E)$. A Factor graph is a bipartite graph $F = (V', E')$ with $V' = V \cup E$ and $E' = \{(s, h) \in V \times E \mid s \in h\}$.

2 Probability distributions on graphs

In order to define a probabilistic graphical model over a graph $G = (V, E)$, each *single vertex* $s \in V$ is associated with a random variabel X_s. The state space of X_s is denoted by $\mathcal{X}_s$ It defines all possible values X_s may take. For example, in the contineous case $\mathcal{X}_s \subseteq \mathbb{R}$ and in the discrete case $\mathcal{X}_s = \{1, \ldots, r\}$. Lower case letters are used to denote an element of the state space, $x_s \in \mathcal{X}_s$. The notation $\{X_s = x_s\}$ corresponds to an event where the random variable X_s takes the specific value $x_s \in \mathcal{X}_s$.

For a *subset of vertices* $A \subset V$, the above notations may be generalized: $X_A := (X_s, s \in A)$. The joint state space is the cartesian product of the individual state spaces, $\mathcal{X}_A := \otimes_{s \in A} \mathcal{X}_s$. A state space element $x_A \in \mathcal{X}_A$ is then given by $x_A = (x_s, s \in A)$. This can be also interpreted as the marginal pdf of $|A|$ random variables with respect to a joint pdf.

2.1 Directed graphical models

Given a directed graph $G = (V, E)$. For an edge $(s \to t)$ s is called *parent* of t and s is called *child* of t. For a given vertex $s \in V$ denote the set of its parents $\pi(t) = \{s \in V | s \to t\}$ (if t has no parents then $\pi(t)$ is the empty set).

- DAG
- interpretation
- not unique (bayes thm)
- plate notation
- model building
- conditional independence

Undirected graphical models

- Markov random fields
- interpretation
- conditional independence
- Moralization

2.2 Factor graphs

Junction tree

•

3 Probabilistc graphical models

In principle there are three kind of graphical representations directed acyclic graphs (DAG), markov networks, factor graphs with addtional types of graphs, which emerge technically upon inference (such as clique graphs and junction trees).

4 DAGs, Markov Networks and Factor Graphs

5 Training

6 Inference

Inference is the process of

7 Basics

- Formal definition of a random variable
- independence and conditional independence

Event x and y are *independent* $:\Leftrightarrow$ their joint distribution factorizes $p(x, y) = p(x)p(y)$ The probability of an event x conditioned on knowing event y is called *conditional probability*, $p(x|y) := \frac{p(x,y)}{p(y)}$ We also say the probability of x given y. If $p(x) = 0$ then $p(x|y)$ is not defined. From this together with $p(x, y) = p(y, x)$ follows Bayes' rule.

Theorem 1 (Bayes' rule)**.**

$$p(x|y) := \frac{p(y|x)p(x)}{p(y)} \tag{1}$$

Indepenent and identically distributed. This is a very common assumption. It is closely related to the notion of symmetry and exchangeable random variables, which eventually leads to the heart of statistical problems: how to characterize joint probabilities. This is turn is closely related to De Fetti's theorem [**?**]. However, for the time beeing we simple state its practical implication. Consider a variable x and n observations $x_1 \dots x_n$ of that variable. They are called independent and identically distributed (IID) $:\Leftrightarrow$ if their joint probability factorizes.

$$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i) \tag{2}$$

Independent refers to the factorization is just the definition above. Identical means that each observation is drawn from the same underlying distribution.

Bayes' Inference. The basic idea is to use Bayes' rule to obtain a distribution over the underlying model parameters of a random process. Let there be given a probability distribution $p(x|\theta)$, which is parametrized by $\theta \in \Theta$ with some parameter space Θ [1]. For a set of n observed data points $X = \{x_1, \dots, x_n\}$ Bayes' rule then reads:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \tag{3}$$

The term $p(\theta|X)$ is called posterior distribution.

Likelihood The term $p(X|\theta)$ in (3) is called likelihood. It describes the probability of the data given the model, which is determined by the (fixed) model parameters θ. How the likelihood is chosen is determined by the underlying random process. A common simplification is to assume IID observations, $p(X|\theta) = p(x_1, \dots, x_n|\theta) = \prod_i p(x_i|\theta)$.

Sometimes it is convenient to work with the logarithm of the likelihood (log-likelihood). In the case of IID observations we get $\sum_i \log p(x_i|\theta)$. The log-likelihood is used if one tries to optimise the likelihood rather than the posterior with respect to θ. Since the log is a strictly monotonic function the optimum remains invariant [2]. The advantage is now that solving for θ is often easier for the log-likelihood than for the likelihood.

[1] Here we are inconsistent in our notation, because capitals usually is reserved for random variables, whereas here it is some parameter space

[2] consider a strictly monotonic, differentiable function g and a differentiable function g. Then $(g \circ f)' = g' \circ f \cdot f'$. Since g is strictly monotonic $g' > 0$. Therefore the maximum of $g \circ f$ is given by the maximum of f.

Maximum likelihood. Instead of working with the full expression 3 when inferring θ one often works directly with the likelihood,

$$\hat{\theta} = \arg \sup_\theta p(X|\theta). \tag{4}$$

This is eventually an manifestation of the likelihood principle [**?**]. Beside the advantages (eg, parametrization invariance, asympotitic properties) the maximum likelihood also has several drawbacks: in high dimensions it can be computationally complex to find the maximum or there could be more local maxima. Furthermore, there is no decision-theoretic and probabilistic support for this approach. In particular, the map $\theta \mapsto p(X|\theta)$ is not a PDF over θ (whereas it is one for $p(\theta|X)$). This is a general property of conditional probabilities.[3]

8 Priors

Prior. The term $p(\theta)$ is called prior distribution. It describes the uncertainty about the model parameter θ before observing (additional) data. So, it reflects our level of ignorance. Determining the prior is the most vodoo-part in the Bayesian approach. Infact it is the main point of criticism. Often one uses uninformative priors, (which are not necessarily uniform priors) or/and priors that are conjugate to a given likelihood function. More details may be found in Ref. [**?**], Chap. 3.

8.1 Conjugate priors

Definition 1 (Conjugacy[**?**])**.** *A family of $\mathcal{F}$ of probability distributions on Θ is called conjugate to a likelihood function $f(x|\theta)$:$\Leftrightarrow \forall \pi \in \mathcal{F}$ the posterior distribution $\pi(\theta|x) \in \mathcal{F}$.*

It is known that for the exponential family there always exists a conjugate prior. Furthermore conjugate priors are quite well studied. Let's summarize the relevant ones for us in Table 1.

Example: Conjugate Prior for the Binomial distribution. This example shows that the Beta distribution is the conjugate prior to the binomial distribution. Assume the model for the data is $P(X|\theta) = Bin(x|\theta, n)$ (θ is

[3]Consider a joint distribution $p(x, \theta)$ with constant marginals, $p(x) = \int p(x, \theta)\, d\theta = 1/k$ on the interval $x = [0, k]$ and $p(\theta) = \int p(x, \theta)\, d\theta = 1/c$ on the interval $\theta = [0, c]$ with $k \neq c$. Assume that the map $\theta \mapsto p(x|\theta)$ is a PDF. Then, $\forall x\ 1 = \int_0^c p(x|\theta)\, d\theta = \int_0^c \frac{p(x,\theta)}{p(\theta)}\, d\theta = c \int_0^c p(x, \theta)\, d\theta = cp(x) = \frac{c}{k} \neq 1$.

Likelihood	Prior	Posterior
$pk(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
$Bin(n,\theta)$	$Beta(\alpha,\beta)$	$Beta(\alpha+x,\beta+n-x)$
$Mult_k(\theta_1\ldots\theta_k,n)$	$Dir_k(\alpha_1\ldots\alpha_k)$	$Dir_k(\alpha_1+x_1\ldots\alpha_k+x_k)$

Table 1: Conjugate priors for given likelihoods.

the success probability and n is the number of trials) and chose the Beta distribution as Prior, $P(\theta) = Beta(\theta|\alpha,\beta)$. Then we get as posterior

$$p(\theta|x,n,\alpha,\beta) = \frac{1}{N} Bin(x|n,\theta) Beta(\theta|\alpha,\beta) \tag{5}$$

$$= \frac{1}{N} \frac{\binom{n}{x}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{B(\alpha,\beta)} \tag{6}$$

$$= Beta(\theta|x+\alpha, n-x+\beta), \tag{7}$$

again a Beta distribution with updated parameters. The factor $1/N$ is just the normalization constant, which may be obtained be integrating the right hand side over θ. The normalization requirement leads to the general trick to only consider the functional form of the free variables (in this case θ) and identify the posterior distribution only via the functional form. The rest is done by the normalization requirement. For this example we would have

$$p(\theta|x,n,\alpha,\beta) \propto Bin(x|n,\theta) Beta(\theta|\alpha,\beta) \tag{8}$$

$$\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \tag{9}$$

$$\propto Beta(\theta|x+\alpha, n-x+\beta) \tag{10}$$

In order to clarify how the conditionals in the first line emerge it's instructive to look at the joint distribution,

$$p(\theta,\alpha,\beta,n,x) = p(x|\theta,n)p(\theta|\alpha,\beta)p(n)p(\alpha)p(\beta)$$

The structure of the underlying model is more transparent in a graph representation (Fig. 1 (a) and (b)). Apart from the Beta and Binomial distributions, there are the additional terms from the individual parameters. These terms eventually cancel in the computation for the conditional distribution,

$$p(\theta|\alpha,\beta,n,x) = \frac{p(\theta,\alpha,\beta,n,x)}{p(\alpha,\beta,n,x)} = \frac{p(\theta,\alpha,\beta,n,x)}{\int p(\theta,\alpha,\beta,n,x)\,d\theta}$$
$$= \frac{p(x|n,\theta)p(\theta|\alpha,\beta)}{\int p(x|\theta,n)p(\theta|\alpha,\beta)\,d\theta},$$

which is just Equation 5 with the explicit form of the normalization constant.

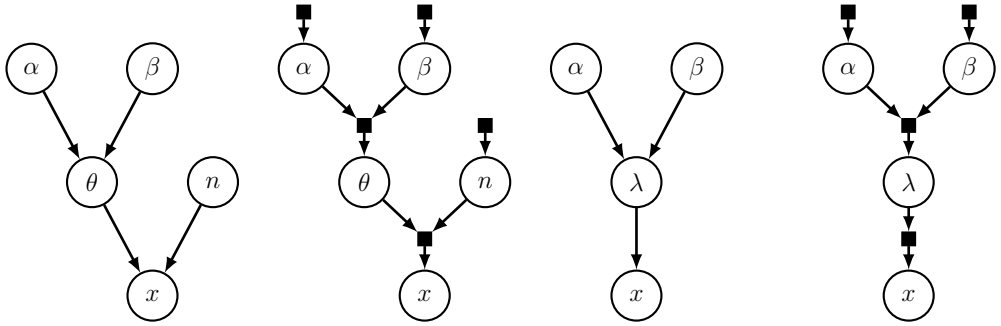

Figure 1: Directed graph (a) and factor graph (b) for beta distribution as conjugate prior of the binomial distribution. (c) and (d) shows the directed and factor graph for the Gamma distribution as conjugate prior of the Poisson distribution.

Example: Conjugate Prior for the Poisson distribution. This example shows that the Gamma distribution is the conjugate prior to the Possion distribution. Assume the model for the data is a Poisson distribution,

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \tag{11}$$

with x number of counts. Chose the Gamma distribution $\Gamma(\lambda|\alpha, \beta)$ as Prior and assume only one Poisson experiment $x = (x)$ (see Fig. **??**. Then we get as posterior

$$\begin{aligned} p(\lambda|x, \alpha, \beta) &\propto \frac{\lambda^x e^{-\lambda}}{x!}\Gamma(\lambda|\alpha, \beta) && (12)\\ &\propto \lambda^{(\alpha+x)-1} e^{-\lambda(\beta+1)} && (13)\\ &\propto \Gamma(\lambda|\alpha + x, \beta + 1) && (14) \end{aligned}$$

So, for one experiment α is updated by the number of counts x and β is updated by the number of experiments (which is one).

Now let's consider the case where we perform n IID Poisson experiments, $x = (x_1, \ldots, x_n)$. **(TODO graph with plate notation)** Then the likelihood becomes

$$p(x|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \tag{15}$$

Then the posterior is obtained by

$$p(\lambda|x,\alpha,\beta) \propto \frac{\lambda^x e^{-\lambda}}{x!}\Gamma(\lambda|\alpha,\beta) \tag{16}$$

$$\propto \lambda^{(\alpha+\sum_{i=1}^n x_i)-1} e^{\lambda(\beta+n)} \tag{17}$$

$$\propto \Gamma(\lambda|\alpha+\sum_{i=1}^n x_i, \beta+n) \tag{18}$$

So, for n Poisson experiments the α parameter of the Gamma distribution is updated by the total number of counts (of all experiments) and beta is updated by the number of experiments (n in this case).

Theorem 2. *There exist conjugate priors for the exponential family of distributions.*

9 Information Theory

Outlook. This should also go into this section

- definition of equivalent parametrization
- Conditional Information
- derivation of Jeffrey's prior
- Flat vs Non-Informative Priors (high dimensional example)

and *p.* [**?**]

Fisher information.

Definition 2 (Fisher information matrix)**.** *Assume that for the likelihood function $f(x|\theta)$, $\theta \in \mathbb{R}^n$ the FI regularity conditions hold [**?**]. The Fisher information matrix is the covariance matrix with respect to the score function*

$$I_{ij}(\theta) := \mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x|\theta) \right],$$

where $\partial/\partial\theta_i \log f(x|\theta)$ is the i-th component of the score function.

The fisher information depends on which of the several possible equivalent parametrizations is chosen.

Definition 3 (Jeffrey's Prior distribution)**.**

$$f(\theta) = c\sqrt{\det I},$$

where I is the Fisher information matrix and c is chosen such that $f(\theta)$ integrates to one if possible.

If no such c exists, then the distribution $f(\theta) = \sqrt{\det I}$ is often used as improper prior distribution [**?**]. In the one dimensional case the fisher information becomes the second moment of the score function and Jeffreys prior is the square root of it. **Check: is the formular below really true (2nd derivative instead of square instead).**

$$I(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial \log f(x|\theta)}{\partial \theta} \right)^2 \right] \tag{19}$$

$$f(\theta) = c\sqrt{I} \tag{20}$$

Example: Binomial distribution. Suppose $X \sim Bin(n, p)$ given $P = p$ for a fixed n. Then, the Fisher information is obtained by

$$\begin{aligned} f_{X|P}(x|p) &= \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n \\ \partial_p (\log f_{X|P}(x|p)) &= \frac{x - np}{p(1-p)} \\ I(p) &= \frac{n}{p(1-p)}, \end{aligned}$$

where we have used that the mean of the Binomial distribution is np and that the variance of the Binomial distribution $\mathbb{E}\left[(x-\mu)^2\right] = np(1-p)$. And Jeffrey's prior becomes,

$$f(p) \propto \sqrt{I} = n\, p^{-\frac{1}{2}} p^{-\frac{1}{2}} \propto Beta\left(1/2, 1/2\right).$$

Therefore, the $Beta(1/2, 1/2)$ distribution is a proper non-informative prior to the Binomial distribution. Note that the Beta distribution is also the conjugate prior to the Binomial distribution.

Example: Poisson distribution. Suppose $X \sim Poi(\lambda)$. The Fisher information is then

$$\begin{aligned} f(x|\lambda) &= \frac{\lambda^x e^{-\lambda}}{x!} \\ \partial_\lambda \log f(x|\lambda) &= \frac{x - \lambda}{\lambda} \\ I(\lambda) &= \frac{1}{\lambda^2} \mathbb{E}\left[(x-\lambda)^2\right] = \frac{1}{\lambda}, \end{aligned}$$

where the formula for variance of the Poisson distribution, $\mathbb{E}\left[(x-\lambda)^2\right] = \lambda$ was used. The Jeffrey's prior is thus,

$$f(\lambda) \propto \sqrt{I} = \lambda^{-\frac{1}{2}} \propto \Gamma(1/2, 0).$$

This seems a bit odd because $\beta = 0$ would lead to the constant zero for the Γ distribution. However, since we are only interested in the proportionality, we can drop the β^α term in the Gamma distribution before setting the values. So, Jeffrey's Prior is an improper Gamma distribution. Improper because $\lambda^{-\frac{1}{1}}$ cannot be normalized. However, it is also of the same form as the conjugate prior. In real application it must be however ensured, that the posterior is a proper distribution. For this case this will be the case as soon as some observations have been made.

In summary the Jeffrey's prior is obtained be requiring invariance under a certain map on the likelihood. It is somewhat against the Baysian mind set where one first chooses a prior on the θ and then uses the likelihood in order to derive the posterior.

Kullback-Leibler divergence. The Kullback-Leibler (KL) divergence measures the 'difference' between two distributions q and p [**?**].This measure of information is designed to measure how far apart two distributions are in the sense of likelihood. That is, if an observation were to come from one of the distributions, how likely is it that you could tell that the observation did not come from the other distribution? [**?**]

Definition 4 (Kullback-Leibler Information)**.**

$$KL(q|p) := \langle \log\, q(x) - \log\, p(x) \rangle_{q(x)} \tag{21}$$

In general $KL(q|p) \geq 0$ and $KL(q|p) = 0 \Leftrightarrow q(x) = p(x)$ (in the sense that the respective probability measures have to be equal). However the KL is not a metric, because $KL(q|p) \neq KL(p|q)$. Even in the symmetrized case (which is sometimes itself called Kl divergence), $KL(q|p) + KL(p|q)$, the triangle inequality does not hold.

10 Linear models

Interpretation

- Predictive interpretation
- Counterfacial interpretation

Interaction Terms

11 Generalized linear models

12 Classical distributions

Binomial distribution, $Bin(n, p)$**.** Probability for the number of successes x for n IID bernoulli trials with chance of success p. The PMF is given by

$$f(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{22}$$

with mean np and variance $np(1-p)$.

Gamma distribution, $\Gamma(\alpha, \beta)$**.** The Γ has a rather generic form and contains the exponential distribution and chi-square distribution as special cases. In econometrics it is frequently used to model waiting times whereas in the Bayes framework it's mainly used as a conjugate prior for rate (inverse scale) parameters (occurring, e.g., in the Poisson or exponential distribution). The PDF of the Gamma distribution is defined as: $x \geq 0$, $\alpha, \beta > 0$

$$f(x|\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-x\beta}}{\Gamma(\alpha)}, \tag{23}$$

where $\Gamma(\alpha)$ is the **Gamma function**. It is the normalizing factor of the distribution to ensure that it integrates to one,

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx \tag{24}$$

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha). \tag{25}$$

The Gamma function can be viewed as a generalization of the factorial to non-integer numbers. That for $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$ can be seen from the recursion formula. The recursion formula is in general very helpful. It can be derived by partial integration, $\partial_\alpha \Gamma(\alpha+1) = \int_0^\infty x^\alpha e^{-x}\, dx = -x^\alpha e^{-x}\Big|_0^\infty + \alpha \int_0^\infty x^{\alpha-1} e^{-x}\, dx = 0 + \alpha\Gamma(\alpha)$.

Note that the Gamma function (24) is indeed the normalization of the Gamma distribution (23). This can be seen by making the substitution $x \mapsto x\beta$ in the integration of Equation (23).

Beta distribution, $Beta(\alpha, \beta)$**.** Is used to model distributions over probabilities because it has a very flexible form. The Beta distribution has the domain of definition $\alpha, \beta > 0, \theta \in [0, 1]$ (some authors have the open interval for theta). The pdf is defined as.

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{26}$$

where B is the **Beta function**. It just normalizes the Beta distribution and it is often more convenient to express it via the Γ function.

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{27}$$

The beta distribution has mean [4] $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. For $\alpha, \beta > 1$ the maximum is given by $\frac{\alpha-1}{\alpha+\beta-2}$.

Poisson distribution, $Poi(\lambda)$. Is used to model the number of events k ('counts') in a fixed (time) interval. The pmf is defined as

$$f(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \mathbb{N}, \lambda \in \mathbb{R}_{>0}, \tag{28}$$

where λ is the event rate or rate parameter. It describes the expected number of events per interval (indeed it is also the expectation value of the Poisson distribution).

Multivariate Normal distribution

$$\mathcal{N}(\mu, \Sigma) = p(x|\mu, \Sigma) \tag{29}$$

$$= -\frac{1}{\sqrt{\det(2\pi\Sigma)}} e^{-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)}, \tag{30}$$

with mean vector μ, covariance matrix Σ and its inverse the precision matrix Σ^{-1}. It can be shown that,

$$\mu = \langle x \rangle_{\mathcal{N}(\mu,\Sigma)} \tag{31}$$

$$\Sigma = \langle (x-\mu)(x-\mu)^\intercal) \rangle_{\mathcal{N}(\mu,\Sigma)} \tag{32}$$

$$\tag{33}$$

[4] To see the mean, we evalute the k-th moment $E(x^k) = \frac{1}{B(\alpha,\beta)} \int x^{\alpha-1+k}(1-x)^{\beta-1} dx$. Multiplying and dividing by $B(\alpha+k, \beta)$ the integrand gets one and we get $E(x^k) = \frac{B(\alpha+k,\beta)}{B(\alpha+k,\beta)}$ The mean is given for $k = 1$. Rewriting in terms of the Γ function we get, $E(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} = \frac{\alpha}{\alpha+\beta}$

Transformations

- $y = Ax$

Let $x \sim \mathcal{N}(\mu, \Sigma)$ and A be a regular Matrix (i.e, non-singular, $\det(A) \neq 0$). Then under the transformtion $y = Ax$, y is again normally distributed, $y \sim \mathcal{N}(A\mu, A\Sigma A^\intercal)$.

Proof. Using the transformation law with jacobian $\det(A)$ gives

$$\begin{aligned} f_Y(y) &= \frac{f_X(A^{-1}y)}{\det(A)} \\ &= \frac{1}{\det(A)\sqrt{2\pi\Sigma}} e^{-\frac{1}{2}(A^{-1}y-\mu)^\intercal \Sigma^{-1}(A^{-1}y-\mu)} \\ &= \frac{1}{\sqrt{2\pi\Sigma_y}} e^{-\frac{1}{2}(y-A\mu)^\intercal \Sigma_y^{-1}(y-A\mu)}, \end{aligned}$$

where we have used $\mu = 1\mu = A^{-1}A\mu$ and identified $\Sigma_y^{-1} = A^{-1\intercal}\Sigma^{-1}A^{-1}$ and therefore $\Sigma_y = A\Sigma A^\intercal$

□

- $z = x + y$

Let $x \sim \mathcal{N}(\mu_x, \Sigma_x)$ and $y \sim \mathcal{N}(\mu_y, \Sigma_y)$ be two independent Normal distributions x. Then $z = x + y$ is again normally distributed with $z \sim \mathcal{N}(\mu_x + \mu_y, \Sigma_x + \Sigma_y)$

Proof. By Theorem **??** we have for the pdf

$$f_{X+Y}(z) = f_X(x) \star f_Y(y) = \int f_X(z-y) f_Y(y)\, dy \tag{34}$$

$$= \frac{1}{\sqrt{\det(2\pi\Sigma_x)}\sqrt{\det(2\pi\Sigma_y)}} \int e^{-\frac{1}{2}\left[(z-y-\mu_x)^\intercal \Sigma_x^{-1}(z-y-\mu_x)+(y-\mu_y)^\intercal \Sigma_y^{-1}(y-\mu_y)\right]} dy \tag{35}$$

We focus on the square bracket in the exponent, define the quantities $\bar{y} = y - \mu_y$, $\bar{z} = z - \mu_x - \mu_y$, and proceed by completing the square with respect to y.

$$[\cdot] = \bar{y}^\intercal \Sigma_y^{-1}\bar{y} + (\bar{y}-\bar{z})^\intercal \Sigma_x^{-1}(\bar{y}-\bar{z}) \tag{36}$$

$$= \bar{y}^\intercal(\Sigma_y^{-1} + \Sigma_x^{-1})\bar{y} - 2\bar{z}^\intercal \Sigma_x^{-1}\bar{y} + \bar{z}^\intercal \Sigma_x^{-1}\bar{z} \tag{37}$$

In order to complete the squere, note that both Σ are symmetric and invertible. Therefore, also the inverse is symmetric and their (inverse) sum is symmetric and invertible. Now, set $\tilde{\Sigma}^{-1} := \Sigma_y^{-1} + \Sigma_x^{-1}$ and consider the term

$$(\bar{y} - \tilde{\Sigma}\Sigma_x^{-1}\bar{z})^\intercal \tilde{\Sigma}^{-1}(\bar{y} - \tilde{\Sigma}\Sigma_x^{-1}\bar{z}) = \bar{y}^\intercal\tilde{\Sigma}^{-1}\bar{y} - 2\bar{z}^\intercal\Sigma_x^{-1}\bar{y} + \bar{z}^\intercal\Sigma_x^{-1}\tilde{\Sigma}\Sigma_x^{-1}\bar{z}$$

Rearranging this identity and plugging it into expression (37), gives

$$[\cdot] = (\bar{y} - \tilde{\Sigma}\Sigma_x^{-1}\bar{z})^\intercal \tilde{\Sigma}^{-1}(\bar{y} - \tilde{\Sigma}\Sigma_x^{-1}\bar{z}) - \bar{z}^\intercal\Sigma_x^{-1}\tilde{\Sigma}\Sigma_x^{-1}\bar{z} + \bar{z}^\intercal\Sigma_x^{-1}\bar{z} \tag{38}$$

$$= (\bar{y} - \tilde{\Sigma}\Sigma_x^{-1}\bar{z})^\intercal \tilde{\Sigma}^{-1}(\bar{y} - \tilde{\Sigma}\Sigma_x^{-1}\bar{z}) + \bar{z}^\intercal(\Sigma_x + \Sigma_y)^{-1}\bar{z}, \tag{39}$$

where in the last step the useful Matrix identities[5] were used. We can now plug this expression back into (35). Unter the integral the first term evaluates to $\sqrt{\det(2\pi\tilde{\Sigma})}, \forall\bar{z}$. Taking into account that $\det\tilde{\Sigma}/(\det\Sigma_x\det\Sigma_y) = 1/\det(\Sigma_x(\Sigma_x^{-1} + \Sigma_y^{-1})\Sigma_y) = 1/\det(\Sigma_x + \Sigma_y)$ eventually gives

$$(35) = \frac{\sqrt{\det(2\pi\tilde{\Sigma})}}{\sqrt{\det(2\pi\Sigma_x)\det(2\pi\Sigma_x)}} e^{-\frac{1}{2}(z-\mu_x-\mu_y)^\intercal(\Sigma_x+\Sigma_y)^{-1}(z-\mu_x-\mu_y)} \tag{42}$$

$$= \frac{1}{\sqrt{\det(2\pi(\Sigma_x + \Sigma_y))}} e^{-\frac{1}{2}(z-\mu_x-\mu_y)^\intercal(\Sigma_x+\Sigma_y)^{-1}(z-\mu_x-\mu_y)} \tag{43}$$

$$= \mathcal{N}(\mu_x + \mu_y, \Sigma_x + \Sigma_y) \tag{44}$$

□

Remark A somewhat less matrix-magic approach ist to pause the computation at equation (38) and realizing that this already establishes a quadratic form in $\bar{z}$ and thus a normal distribution in z. In order to get the expressions vor μ_z and Σ_z one could alternatively calculate them by using the linearity property of the expectation value together with the independence assumption $f_{X,Y}(x,y) = f_X(x)f_Y(y)$:

$$\mu_z = \langle z\rangle = \iint f_X(x)f_Y(y)(x+y)\,dx\,dy = \langle x\rangle + \langle y\rangle = \mu_x + \mu_y \tag{45}$$

[5]Let X, Y be regualar matrices. Then:

$$X^{-1}(X^{-1} + Y^{-1})^{-1}Y^{-1} = (X+Y)^{-1}. \tag{40}$$

Set the left hand side $Z := X^{-1}(X^{-1}+Y^{-1})^{-1}Y^{-1}$. Its inverse is apparently $Z^{-1} = X+Y$ Therefore, $Z = (X+Y)^{-1}$. The second matrix identity reads

$$X^{-1} - X^{-1}(X^{-1} + Y^{-1})^{-1}X^{-1} = (X+Y)^{-1}. \tag{41}$$

The left hand side can be rewritten as $X^{-1} - X^{-1}(X^{-1} + Y^{-1})^{-1}(X^{-1} \pm Y^{-1}) = X^{-1}(X^{-1} + Y^{-1})^{-1}Y^{-1} = (X+Y)^{-1}$, where in the last step identity (40) was used.

$$\Sigma_z = \langle (x+y-\mu_x-\mu_y)(x+y-\mu_x-\mu_y)^\intercal \rangle \tag{46}$$
$$= \langle (x-\mu_x)(x-\mu_x)^\intercal \rangle + \langle (y-\mu_y)(y-\mu_y)^\intercal \rangle + 2\langle (x-\mu_x)(y-\mu_y)^\intercal \rangle \tag{47}$$
$$= \langle (x-\mu_x)(x-\mu_x)^\intercal \rangle + \langle (y-\mu_y)(y-\mu_y)^\intercal \rangle, \tag{48}$$

where the symmetry of Σ was used (factor 2) and the last term vanishes because of independence ($\iint f_X(x) f_Y(y)(x-\mu_x)(x-\mu_x)^\intercal = 0$).

- $z = Ax + y$

Let $x \sim \mathcal{N}(\mu_x, \Sigma_x)$ and $y \sim \mathcal{N}(\mu_y, \Sigma_y)$ be two independent Normal distributions x. Then $z = Ax + y$ with A being a regular matrix is again normally distributed with $z \sim \mathcal{N}(A\mu_x + \mu_y, A\Sigma_x A^\intercal + \Sigma_x)$

Proof. This follows immeadiatly by composition of the above results. Set $\tilde{x} = Ax$. Then $z = \tilde{x} + y \sim \mathcal{N}(\mu_{\tilde{x}} + \mu_y, \Sigma_{\tilde{x}} + \Sigma_y) = \mathcal{N}(A\mu_x + \mu_y, A\Sigma_x A^\intercal + \Sigma_y)$ □

13 Variational Inference

Variational inference [4, 1] is an alternative to MCMC in the Bayes world. In essence, the posterior distributions are approximated by optimizing over a family of testfunctions on a scalar bound, the KL diverence. In the special case where we are interested in the point estimate of of model parameters, this gets more structure and leads [13] to the EM algorithm [5, 8](which also gives an estimate for the hidden variables). [6].

For many practical applications the mean field approximation over the variational densities is applied which, leads to coordinate ascent mean-field variational inference (CAVI) [3]. CAVI can be seen as "message passing" algorithm and thus connects variational inference with graphical models [16, 15, 9, 12] (implementation https://dotnet.github.io/infer/).

From a numerical point of view there have been stochastic generalizations in order to deal with large datasets. In order to avoid computing the functional derivations explicitly for a given model analytically an "automatic differentiation variational inference" (ADVI) method has been proposed [10].

The problem. Assume that we have random variables X of observed data and Z of latent random variables. Inference of the model $p(x, z) = p(x|z)p(z)$ amounts to finding the conditional probabilities $p(z|x)$. This quantity is also known posterior. Formally this is trivial via Bayes theorem $p(z|x) = p(x, y)/p(x)$. However computing $p(x) = \int p(x, z)dz$ may be computationally intractable (for example, it may lead to very high dimensional numeric

integrations for correlated latent variables). Cynically, $p(x)$ is just a normalization factor. In physics it is called partition function and in statistics its called evidence. In fact, computing the evidence is one of the most challenging parts of Bayesian methods.

The optimization problem. Variational inference tries to approximately determine $p(z|x)$ by introducing a family of approximate probability densities $q(z|x) \in \mathcal{Q}$ and then finding that member that approximates the posterior optimally with respect to the KL divergence,

$$q^*(z) = \underset{q(z)\in\mathcal{Q}}{\operatorname{argmin}} \, D(q(z) \parallel p(z|x)), \tag{49}$$

where the shorthand notation $q(z) \equiv q(z|x)$ was introduced. This transforms an integration problem to an optimization problem. However, the KL divergence of this form is of no help to the original problem (computing the marginal $p(x)$) as we still need to compute the term $p(z|x)$ which, requires the term $p(x)$. So we need to find another objective function, the evidence lower bound (ELBO),

$$\mathcal{F}(q) := \langle \log p(x,z) \rangle_{q(z)} - \langle \log q(z) \rangle_{q(z)} \tag{50}$$

Theorem 3. *Let there be given a family of probability densities $\mathcal{Q}$. Then $q^*(z)$ is a minimizer over $\mathcal{Q}$ of $D(q(z) \parallel p(z|x))$ iff it is a maximizer of the ELBO $\mathcal{F}(q)$.*

Proof. This follows from

$$\begin{aligned} D(q(z) \parallel p(z|x)) &= \langle \log q(z) \rangle_{q(z)} - \langle \log p(z|x) \rangle_{q(z)} \\ &= -\mathcal{F}(q) + \langle \log p(x) \rangle_{q(z)} \\ &= -\mathcal{F}(q) + p(x), \end{aligned}$$

and that $p(x)$ is constant with respect to $q(z)$, $\delta p(x)/\delta q(z) = 0$ □

So, first we need to specify a family of variational densities $\mathcal{Q}$ and then optimize Equation 50 over $\mathcal{Q}$. Not that both, the KL and ELBO objective functions implicitly depend on x and so does $q^*(z)$. This just reflects the dependence on the training data.

Discussion. Turning to the interpretation of variational inference rewrite

$$\begin{aligned} \mathcal{F}(q) &= \langle \log p(x|z) \rangle_{q(z)} + \langle \log p(z) \rangle_{q(z)} - \langle \log q(z) \rangle_{q(z)} \\ &= \langle \log p(x|z) \rangle_{q(z)} - D(q(z) \parallel p(z)). \end{aligned} \tag{51}$$

The density that maximizes the ELBO thus compromises between i) putting most probability weight on z where the posterior $p(x|z)$ is large and ii) beeing close to the prior $p(z)$. Furthermore if we assume that the N observed data points are iid of the form $x = x_1, \ldots, x_N$ the first expression reads $\langle \log p(x|z) \rangle_{q(z)} = \sum_{i=1}^{N} \langle \log p(x_i|z) \rangle_{q(z)}$. Thus, the first term becomes more important for large number of observed data points. But this is just in line with the usual Bayesian behavior.

Furthermore the ELBO bounds the (log) evidence $p(x)$,

$$\log p(x) = \mathcal{F}(q) + D(q(z) \parallel p(z|x)) \geq \mathcal{F}(q) \tag{52}$$

because $D(\cdot) \geq 0$. This property lends ELBO its name and has been used for model selection under the assumption that the ELBO is a good approximation to the marginal likelihood. However, this approach is not rigorous. So we will just use the ELBO as objective function for the optimization problem (which is rigorous).

The KL-divergence enjoys the property $D(q \parallel p) = 0 \Leftrightarrow q = p$. Assume a maximizer q^* of the ELBO, which is also a minimizer of the KL-divergence. Now, if the KL-divergence $D(q^*(z) \parallel p(z|l)) = 0$ then q^* is the true posterior. However this need not be the case as the variational family typically suffers from severe restrictions due computational resources. In this case it is not clear how "good" the minimizer of the KL-divergence approximates the posterior [12, 11].

13.1 Mean-field approximation

Definition 5 (Mean field approximation)**.** *A family of pdfs $q_Z(z) \in \mathcal{Q}$ is called mean field approximation if it is of the form $q_Z(z) = \prod_\ell q_{Z_\ell}(z_\ell)$ with $q_{Z_\ell}(z_\ell) \in \mathcal{Q}_\ell$ and $Q = \otimes_\ell \mathcal{Q}_\ell$.*

Note that the definition does not assume how the latent variables are factorized. The case where the latent variables are fully factorized is sometimes called *naive mean field approximation* while the case where some latent variables are collected into groups is sometimes called *generalized mean field approximation*. In the following we use the shorthand notation $q_\ell(z_\ell) \equiv q_{Z_\ell}(z_\ell)$

Theorem 4. *Given a mean field approximation of a variational family, $\prod_\ell q_\ell(z_\ell) \in \otimes_\ell \mathcal{Q}_\ell$ with $\mathcal{Q}_\ell = \{q_\ell(z_\ell) \,|\, q_\ell(z_\ell)$ is pdf$\}$. Then the maximizer of the ELBO is given by*

$$q^*(z) = \prod_\ell q_\ell^*(z_\ell), \tag{53}$$

$$q^*(z_\ell) = \frac{1}{N_\ell} \exp \langle \log p(x, z_\ell, z_{-\ell}) \rangle_{q_{z_{-\ell}}}, \tag{54}$$

where $z = \{z_\ell, z_{-\ell}\}$, $z_{-\ell} = \{z_1, \dots, \cancel{z_\ell} \dots\}$ *and* $q_{z_{-\ell}} = \prod_{1 \neq \ell} q_i(z_i)$

Proof. Taking the functional derivative of the ELBO under the constraint that all $q_\ell(z_\ell)$ are pdfs, i.e., $\int q_\ell(z_\ell)\, dz_\ell = 1$ and $q_\ell(z_\ell) \geq 0\ \forall z_\ell$ via Lagrangian multipliers gives,

$$
\begin{aligned}
&\frac{\delta}{\delta q_k(z'_k)} \left\{ \mathcal{F}(\prod_{\ell=1}^{L} q^*_\ell(z_\ell)) - \sum_{\ell=1}^{L} \lambda_\ell \left(\int q_\ell(z_\ell)\, dz_\ell - 1 \right) \right\} \\
=&\frac{\delta}{\delta q_k(z'_k)} \left\{ \int dz_1 \cdots \int dz_1\, p(x, z_1, \dots, z_L) \prod_{\ell=1}^{L} q_\ell(z_\ell) - \sum_{\ell=1}^{L} \int dz_\ell\, q_\ell(z_\ell) \log q_\ell(z_\ell) \right\} \\
&- \lambda_k \int \delta(z_k - z'_k)\, dz_k \\
=&\int \cdots \int dz_1 \dots \cancel{dz_k} \dots dz_L\, p(x, z_1, \dots, z'_k, \dots z_L) \prod_{\ell \neq k} q_\ell(z_\ell) - \log q_k(z'_k) - 1 - \lambda_k.
\end{aligned}
$$

Solving for $q_k(z'_k)$ and using the notation from above gives

$$
q_k(z'_k) = \frac{\exp\langle \log p(x, z'_k, z_{-k}) \rangle_{q_{z_{-k}}}}{\exp(1 + \lambda_k)} \tag{55}
$$

The Lagrangian multiplier and thus the is obtained by normalization, $N_k = \exp(1 + \lambda_k) = \int dz'_k\, \exp\langle \log p(x, z'_k, z_{-k}) \rangle_{q_{z_{-k}}}$. Together with the functional from this implies that q_k is indeed a pdf. □

Note that (54) can be equivalently written in terms of complete conditionals $p(z_\ell | x, z_{-\ell})$,

$$
q^*(z_\ell) = \frac{1}{N'_\ell} \exp \langle \log p(z_\ell | x, z_{-\ell}) \rangle_{q_{z_{-\ell}}}, \tag{56}
$$

because $\langle \log p(x, z_\ell, z_{-\ell}) \rangle_{q_{z_{-\ell}}} = \langle \log p(z_\ell | x, z_{-\ell}) \rangle_{q_{z_{-\ell}}} - \langle \log p(x, z_{-\ell}) \rangle_{q_{z_{-\ell}}}$ and the second term does not depend on $q_\ell(z_\ell)$ and thus may be absorbed into the normalization factor.

- Theorem 3 and 4 establish the best approximation to the KL-divergence of the true posterior within a factorized family of variational densities. Recall $q(z) \equiv q(z|x)$ and $z = \{z_1, \dots, z_\ell, \dots\}$
- Integrating out definition 5 shows that the approximation to the marginal posterior corresponds to the variational component, which is also complete conditional.

$$
q(z_\ell | x) \equiv q(z_\ell) = q_\ell(z_\ell) \tag{57}
$$

$$
q(z_\ell | z_{-\ell}, x) = q_\ell(z_\ell) \tag{58}
$$

Algorithm 1: Coordinate ascent variational inference

Input: A model $p(x, z)$ and a dataset x
Result: Maximizer of the ELBO
Initialize: variational densities $q_\ell(z_\ell),\ \ell = 1 \dots m$
$n = 0$
while *While the ELBO is not converged* **do**
  for $\ell = 1 \dots m$ **do**
    $q^{(n+1)}(z_\ell) \propto \exp \langle \log p(x, z_\ell, z_{-\ell}) \rangle_{q^{(n+1)}_{z_{<\ell}} \cdot q^{(n)}_{z_{>\ell}}}$
  end
  Compute ELBO $\mathcal{F}(q^{(n+1)})$ from Eq. (50)
  $n = n + 1$
end
return $q^{(n)}(z) = \prod_{\ell=1}^{m} q_\ell^{(n)}(z_\ell)$

- From Theorem 4, $q^*_\ell(z_\ell)$ depends on all other approximate pdfs of random variables. In practice this yields an iterative scheme of computation, the *coordinate ascent mean-field variational inference algorithm* (CAVI)[4, 3]. This fix-point procedure is summarized in Algorithm 1. The short-hand notation that $q^{(n+1)}_{z_{<\ell}} \cdot q^{(n)}_{z_{>\ell}} \equiv q_1^{(n+1)} \dots q_{\ell-1}^{(n+1)} q_\ell^{(n)} \dots q_\ell^{(n)}$ was used to indicate that the updated variational densities are used as soon as they are available [3].

- Especially from Eq (56) the connection between CAVI and the Gibbs sampler in the MCMC approach [6, 7] can be seen. Both use the complete conditional $p(z_\ell|x, z_{-\ell})$ in their update steps.

- In general this formulation is not very useful for practical computations. Thus typically further structure is incorporated by using the exponential family as variational family, which leads to much simpler CAVI update equations and allows variational inference to scale to massive data [4, 2, 14].

Variational Bayesian Expectation-Maximization (VB-EM). So far there was no interpretation on the latent variables. Consider a model with latent variables $z = \{z, \theta\}$, where θ are model parameters and z are hidden (e.g., local) variables. The Variational Bayesian Exptectation-Maximisation (VB-EM) [1, 2] is just a consequence of Theorem 4 together with the mean-

field ansatz $q(z, \theta) = q_z(z)q_\theta(\theta)$,

$$q_z(z) \propto \exp \langle \log p(z, x, \theta) \rangle_{q_\theta}, \tag{59}$$
$$q_\theta(\theta) \propto \exp \langle \log p(x, z, \theta) \rangle_{q_z}, \tag{60}$$

which can be equivalently written as

$$q_z(z) \propto \exp \langle \log p(z, x|\theta) \rangle_{q_\theta}, \tag{61}$$
$$q_\theta(\theta) \propto p(\theta) \exp \langle \log p(x, z|\theta) \rangle_{q_z}. \tag{62}$$

This follow $p(x, z, \theta) = p(x, z|\theta)p(\theta)$. The residual contribution $\exp \langle \log p(\theta) \rangle_{q_\theta}$ may be absorbed in the normalization factor in (61) and $\exp \langle \log p(\theta) \rangle_{q_z} = p(\theta)$ in (62).

13.2 Parametric variational distributions

13.3 Point estimates

Very often θ is the only quantity of interest, either 1) as a marginal posterior pdf, $\int p(\theta, z|x)\, dz$, or 2) by a point-estimate of it (MAP) $\operatorname{argmax}_\theta \int p(\theta, z|x)\, dz$, or 3) as a maximum likelihood estimate (MLE) $\operatorname{argmax}_\theta \int p(x, z|\theta)\, dz$. Clearly, the first case can be approximately obtained by VB-EM or equivalently, the mean-field ansatz $q(z, \theta) = q(z)q(\theta)$ and using the machinery of Theorem 4.

Expectation-Maximisation The aim of the expectation maximisation (EM) algorithm is provide a maximum likelihood estimate (MLE) of the model parameters θ under hidden random variables Z, $\operatorname{argmax}_\theta p(x|\theta) = \operatorname{argmax}_\theta \int p(x, z|\theta)\, dz$.

Theorem 5. *Let there be given a variational family $q \in \mathcal{Q}$. Then*

$$\mathcal{F}(q, \theta) := \langle \log p(x, z|\theta) \rangle_{q(z)} - \langle \log q(z) \rangle_{q(z)} \tag{63}$$

is a lower bound to the log-likelihood $\log p(x|\theta) \geq \mathcal{F}(q, \theta)\ \forall q$ with equality iff $q(z) = p(z|x, \theta) \in \mathcal{Q}$.

Proof. We have $\log p(x|\theta) = \mathcal{F}(q, \theta) + D(q(z) \parallel p(z|x, \theta)) \geq \mathcal{F}(q, \theta)$. Equality follows from the property of the KL-divergence $D(q \parallel p) = 0 \Leftrightarrow q = p$ □

Note that the variational densities only depend on the observed data $q(z) \equiv q(z|x)$ (but not on parameter θ). In contrast to the fully Baysian variational methods above we need the bounding property of theorem 5 in order to lend the maximizer of $\mathcal{F}(q, \theta)$ an interpretation as beeing the MLE.

Algorithm 2: Classical EM Algorithm (n-th update)

E-step: $q^{(n)}(z) = p(z|x, \theta^{(n-1)})$
M-step: $\theta^{(n)}(z) = \text{argmax}_\theta \langle \log p(x, z|\theta) \rangle_{q^{(n)}}$

For a given variational family $\mathcal{Q}$ we can now formulate the *constraint* (with respect to $\mathcal{Q}$) EM Algorithm [1] by optimizing functional (63):

$$q^*(z) = \text{argmax}_q \mathcal{F}(q, \theta) \quad \text{(E-step)} \tag{64}$$

$$\theta^* = \text{argmax}_\theta \mathcal{F}(q, \theta) \quad \text{(M-step)}, \tag{65}$$

where $\text{argmax}_\theta \mathcal{F}(q, \theta) = \text{argmax}_\theta \langle \log p(x, z|\theta) \rangle_q$ because $q(z)$ does not depend on θ.

If the varitional family is sufficiently large such that $= p(z|x, \theta) \in \mathcal{Q}$ (in other words if the conditional posterior $p(z|x, \theta)$ is computable) the the equality condition of theorem 5 can be used to arrive at the *unconstraint*, or classical EM algorithm. [5, 13]

$$q^*(z) = p(z|x, \theta) \quad \text{(E-step)} \tag{66}$$

$$\theta^* = \text{argmax}_\theta \langle \log p(x, z|\theta) \rangle_q \quad \text{(M-step)}. \tag{67}$$

The M-step follows by taking the functional derivative with respect to $q(z)$ under the constraint $\int q(z)\, dz = 1$. While the constraint EM converges to lower bound of the log-likelihood the uncontraint converges to a local maximum of the log-likelihood. Both variants lead to an iterative algorithm (Algorithm 2).

Theorem 5 and the (un)constraint EM formulas are only statements on what happens at the maximizer but not *how* the maximizer is computationally approached. This allows to speed up algorithm 2 for large data sets. Consider a case with global model parameters θ and local hidden random variables $Z = \{Z1, \dots Z_N\}$ together with N observations $X = \{X_1 \dots X_N\}$ such that $p(x, z|\theta) = \prod p(x_i, z_i|\theta)$. Then choose $q(z) = \prod q_i(z_i)$. If the goal is to learn θ there is no need to compute the E-step for all datapoints at each iteration. It is enough to perform the E-step for only one datapoint component as this increases $\mathcal{F}(q, \theta)$ for fixed θ (algorithm 3). Eventually this leads to "sparse" algorithms which have been shown to converge faster **??**.

References

[1] Matthew J. Beal. Variational algorithms for approximate bayesian inference. Technical report, 2003.

Algorithm 3: Sparse EM Algorithm (n-th update)

Choose some data point i to be updated

$$\text{E-step:} \quad q_k^{(n)}(z) = \begin{cases} p(z_i|x_i, \theta^{(n-1)}) & k = i \\ q_k^{(n)}(z) & k \neq i \end{cases}$$

$$q^{(n)}(z) = \prod_k q_k^{(n)}(z)$$

$$\text{M-step:} \quad \theta^{(n)}(z) = \text{argmax}_\theta \langle \log p(x, z|\theta) \rangle_{q^{(n)}}$$

[2] J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, M. West (eds, Matthew J. Beal, and Zoubin Ghahramani. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures, 2003.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians, 2016.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[6] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[7] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, November 1984.

[8] Maya R. Gupta and Yihua Chen. Theory and use of the em algorithm. *Found. Trends Signal Process.*, 4(3):223–296, March 2011.

[9] David A. Knowles and Tom Minka. Non-conjugate variational message passing for multinomial and binary regression. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1701–1709. Curran Associates, Inc., 2011.

[10] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference, 2016.

[11] M.A.R. Leisink and H.J. Kappen. A tighter bound for graphical models, 2001.

[12] Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, January 2005.

[13] Radford M. Neal and Geoffrey E. Hinton. Learning in graphical models. chapter A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants, pages 355–368. MIT Press, Cambridge, MA, USA, 1999.

[14] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[15] Matthew Wand, John Ormerod, Simone Padoan, and Rudolf Fuhrwirth. Mean field variational bayes for elaborate distributions. *Bayesian Analysis*, 6, 12 2011.

[16] John Winn and Christopher M. Bishop. Variational message passing. *J. Mach. Learn. Res.*, 6:661–694, December 2005.