An Introduction to Bayesian Neural Network and Uncertainty Quantification in Deep Learning SUPPLEMENTARY MATERIAL

Inverted CERN School of Computing 2023

Jacopo Talpini

Outline

On Marginalization

- As we have seen BNNs rely on marginalizing the posterior.
- Marginalization is not controversial, it is an application of the sum and product rules of probability theory. For instance, the pdf for two random variables a, b can always be factorized as:

$$p(a, b) = p(a|b)p(b) = p(a)p(b|a) \tag{1}$$

$$p(a, b|c) = p(a|b, c)p(b|c) = p(a|c)p(b|a, c) \tag{2}$$

- Given $p(a, b)$ we can project it onto the coordinate axes and obtain the marginal distributions for a and b:

$$p(a) = \int p(a, b)db = \int p(a|b)p(b)db \tag{3}$$

$$p(b) = \int p(a, b)da = \int p(b|a)p(a)db \tag{4}$$

On Marginalization

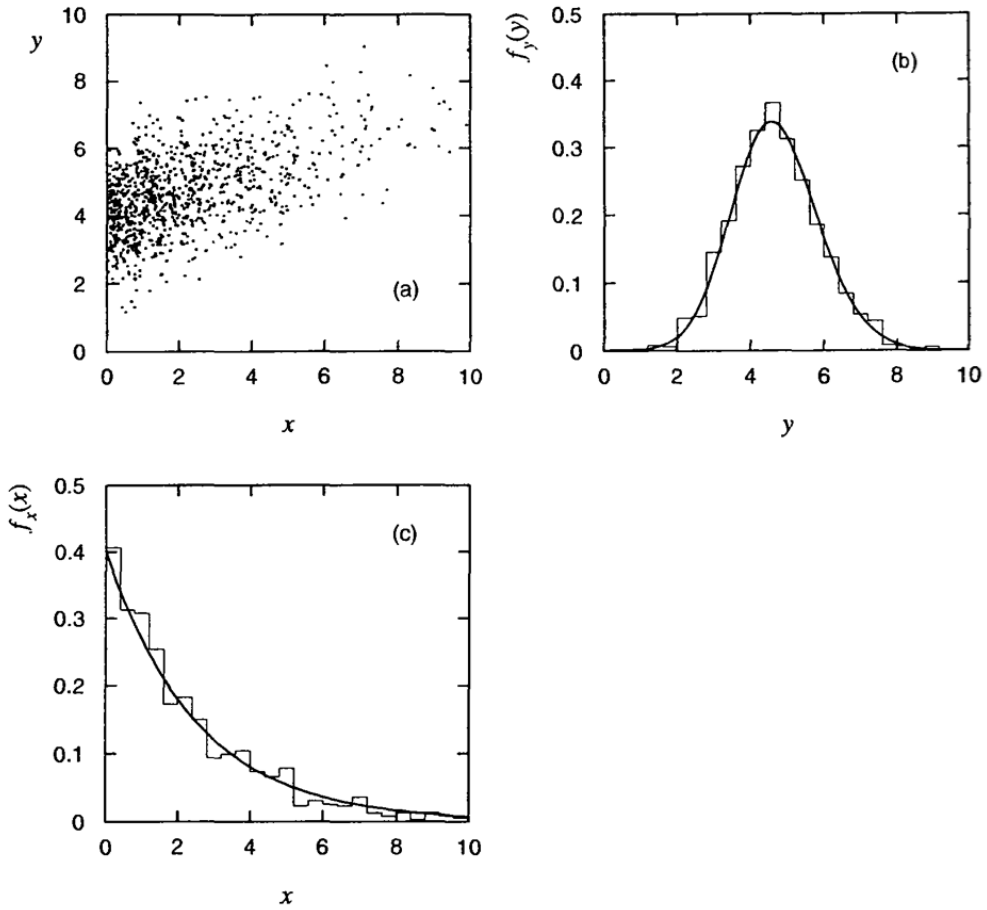

Fig. 1.5 (a) The density of points on the scatter plot is given by the joint p.d.f. $f(x, y)$. (b) Normalized histogram from projecting the points onto the y axis with the corresponding marginal p.d.f. $f_y(y)$. (c) Projection onto the x axis giving $f_x(x)$.

1

[1] From: Cowan, Glen. Statistical data analysis. Oxford university press, 1998.

On Marginalization: Inference

- In Bayesian inference it is common to marginalize away variables we want to get rid of, like nuisance parameters, trough Eq.(3) .
  In fact, a Bayesian treats parameters of models just like any other unknown random variable, and applies the rules of probability theory to infer them from data.
- Remember that in a frequentist setting we are not allowed to think about the notion of probability distributions as applied to parameter sets. Inference gives point estimates for the paramaters, based on the likelihood function which is NOT a pdf for the parameters, given the data, so Eq.(3) simply does not hold in this setting.
- Marginalization comes at a price: we have to specify a prior and the computation is non trivial

On Marginalization: MCMC

- Markov Chain Monte Carlo (MCMC) are methods for sampling pdf even if we do NOT have a full analytical description of the properly normalized pdf. This is why MCMC are ideal for sampling posterior pdf.
- Given a likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$ we can obtain the posterior pdf :

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{5}$$

- where $Z = p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the evidence and typically it is extremely hard/impossible to calculate.
- Given the posterior we may be interested in computing the mean or marginalize, i.e. in general computing :

$$E_{p(\theta)}[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{6}$$

On Marginalization: MCMC

- A good sampling technique allows to approximate well the integral in Eq.(6):

$$E_{p(\theta)}[g(\boldsymbol{\theta})] \sim \frac{1}{K}\sum_{k=1}^{K} g(\boldsymbol{\theta}_k) \tag{7}$$

- Where $\{\boldsymbol{\theta}_k\}_{k=1}^{K}$ are K sampled points from the pdf $p(\boldsymbol{\theta})$
- As previously pointed out, $p(\boldsymbol{\theta})$ (e.g. a posterior) is known up to a normalizing factor. The beauty of MCMC is that sampling and hence Eq.(7) can be computed without the need of knowing the normalizing constant Z.
- Moreover, MCMC allows to observe the distribution of any single component or set of components of the vector $\boldsymbol{\theta}$, (i.e. marginalizing over the other components) trivially, by projecting the sampling to the subspace of interest.

Outline

Variational Inference

- Another possibility when dealing with a nasty pdf is Variational Inference
- Idea: approximate the nasty pdf $p(\boldsymbol{\theta}) = (1/Z)p^*(\boldsymbol{\theta})$ with a more tractable (=easy to sample) pdf $Q(\boldsymbol{\theta}|\boldsymbol{\phi})$ and adjust ϕ to get the best approximation. Then:

$$E_{p(\theta)}[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \sim \int g(\boldsymbol{\theta})Q(\boldsymbol{\theta}|\boldsymbol{\phi})d\boldsymbol{\theta} \tag{8}$$

- We have to define the goodness of the aproximation, the standard approach is based on the KL-divergence. A possible issue is that it is not symmetric, so we have:

$$KL(p||Q) = \int p(\boldsymbol{\theta})\log\left(\frac{p(\boldsymbol{\theta})}{Q(\boldsymbol{\theta}|\boldsymbol{\phi})}\right)d\boldsymbol{\theta} \tag{9}$$

$$KL(Q||p) = \int Q(\boldsymbol{\theta}|\boldsymbol{\phi})\log\left(\frac{Q(\boldsymbol{\theta}|\boldsymbol{\phi})}{p(\boldsymbol{\theta})}\right)d\boldsymbol{\theta} \tag{10}$$

On KL-Divergence

- Both are minimized when Q is close to p but their behaviour is different
- For example let $p^* = [1/3, 1/3, 1/3, \epsilon]$ and $Q = [1/4, 1/4, 1/4, 1/4]$. In this case:
  $KL(p||Q) \sim log(4/3)$
  $KL(Q||p) = (3/4)log(3/4) + (1/4)log(1/4\epsilon)$
- $KL(Q||p)$ heavily penalize configurations in which Q puts high density where p is close to zero
- $KL(p||Q)$ tends to give less compact, broader approximations of p

On KL-Divergence

- Moreover, we have to keep in mind that p is supposed to be a nasty distribution, so it might seem unlikely that we are able to compute easily $KL(p||Q)$ which involves an expectation w.r.t. p
- On the other hand $KL(Q|p)$ can be computed easily (as long as Q is tractable) and it is possible to show (see the lecture) that the ELBO $\mathcal{F}(\theta)$ does not depends at all on the nasty p, and it is lower bounded by the log-evidence.

Outline

TO DO