Lecture 2: An introduction to stochastic variational inference

luca.ambrogioni

December 2020

Exact Bayesian inference is possible only in models with conjugate priors. However, in most machine learning problems the likelihood is not conjugate to the prior and the inference cannot be carried out in closed form. In the context of this course, we will often use very complex and highly parameterized deep learning components for both priors and likelihoods. In these settings, exact inference is never possible and we have to rely on approximate inference methods. Traditionally, approximate Bayesian inference has been performed with sampling methods such as MCMC. However, sampling is often computationally inefficient and it is difficult to use successfully in deep models. Therefore, in this course will use stochastic gradient-based variational inference which reduces inference to a non-convex optimization problem that can be solved using standard deep learning methods.

0.0.1 Discretization of the latent space and the curse of dimensionality

We start by discussing grid integration, the simplest approximate inference method. This method is usually useless in probabilistic deep learning but it is useful in order to understand the challenges behind approximate inference.

Consider a Bayesian inference problem where z is a latent variable and $D = \{x_1, \dots, x_N\}$ is a set of observations. Bayes theorem allows you to obtain the posterior distribution over the latent given a prior distribution and a likelihood model:

$$p(z \mid D) = \frac{p(D \mid z)p(z)}{p(x)} \tag{1}$$

where $p(x)$ is the likelihood of the data given the model where the latent variable z has been marginalized out:

$$p(D) = \int p(D \mid z)p(z)\mathrm{d}z \ . \tag{2}$$

The integral in Eq. 0.0.1 can be solved analytically only when the prior is conjugate to the likelihood. In all other cases, the integral can only be be evaluated numerically using approximate methods. If the dimensionality of

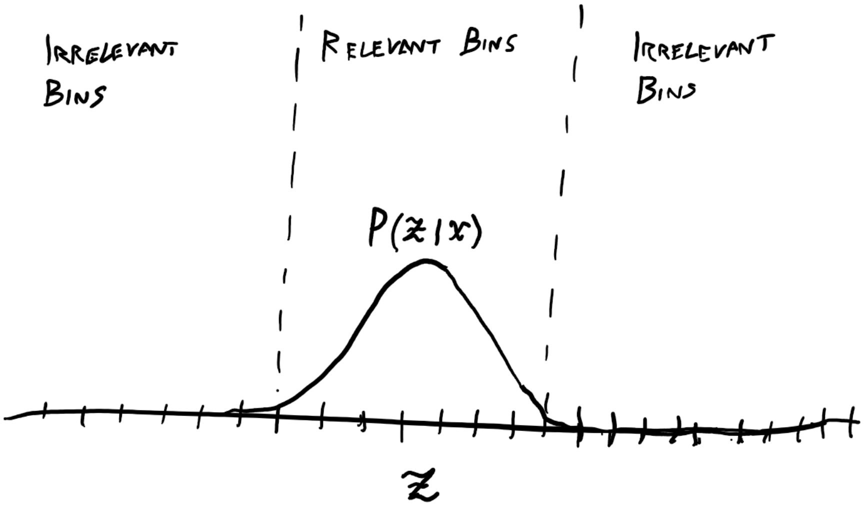

Figure 1: Relevant and irrelevant bins.

z is low (up to 4-5), we can solve the integral numerically using a numerical integration method. For example, if z is one-dimensional we can approximate the integral by discretizing the range of z into N bins and computing a finite sum:

$$\int p(D \mid z)p(z)\mathrm{d}z \approx \sum_{n=1}^{N} p(D \mid z_n)p(z_n)\Delta z \;, \tag{3}$$

where Δz is the bin size. Unfortunately this approach becomes quickly unfeasible when the dimensionality of z increases as the number of bins scale as N^D and therefore increase exponentially with the dimensionality. This is referred in the literature as the *curse of dimensionality.* Superficially, this problem seems to be insurmountable and in fact Bayesian inference is intractable in the general case. However, fortunately the mass of the posterior probability $p(z \mid x)$ usually concentrate in a very small fraction of those bins with all the other ones giving a negligible contribution. The division between relevant and irrelevant bins is visualized in Fig. **??** for the 1D case. Importantly, when the dimensionality is high the number of relevant bins is usually a very small fraction of the total number, this implies that we can perform reliable approximate inference if we are able to localize these relevant bins and exclude everything else.

0.0.2 Importance sampling

We can now discuss importance sampling, a very simple but inefficient sampling method that will prepare the ground for the variational approach. As we saw in the previous section, in high-dimensional problems we only have hope to obtain a good approximation of the model evidence if we restrict our attention to the

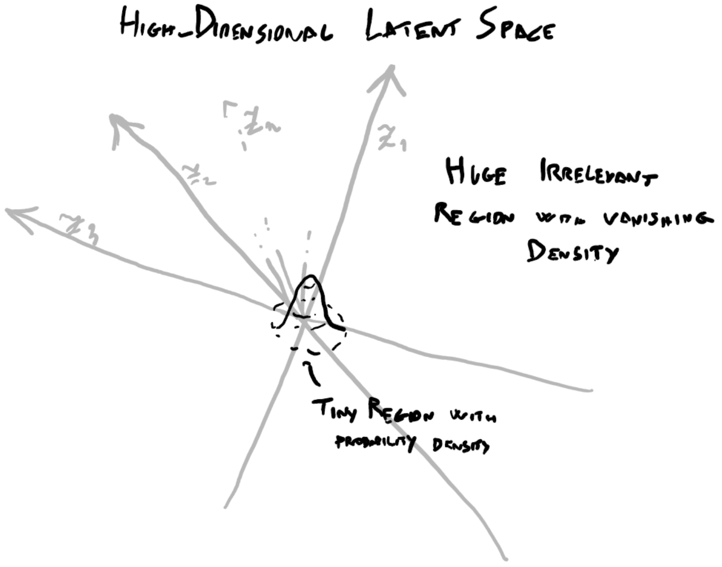
HIGH-DIMENSIONAL LATENT SPACE
z_n
z_1
z_2
z_3
HUGE IRRELEVANT REGION WITH VANISHING DENSITY
TINY REGION WITH PROBABILITY DENSITY

Figure 2: Relevant and irrelevant regions in a high dimensional latent space.

(hopefully small) fraction of the latent space with a non-negligible amount of probability density. Let's assume for now that we have a good guess of where the probability density of the posterior is concentrated. How can we exploit this information? The most straightforward approach is to introduce a new sampling distribution $q(z)$ that is concentrated in this high density region and to rewrite the model evidence as an expectation over this sampling distribution:

$$
\begin{aligned}
p(D) &= \int p(D \mid z)p(z)\mathrm{d}z \ , && (4)\\
&= \int p(D \mid z)p(z)\frac{q(z)}{q(z)}\mathrm{d}z \ , && (5)\\
&= \int p(D \mid z)\frac{p(z)}{q(z)}q(z)\mathrm{d}z \ , && (6)\\
&= \mathbb{E}_{z\sim q(z)}\left[p(D \mid z)\frac{p(z)}{q(z)}\right] \ . && (7)
\end{aligned}
$$

In practice, we can approximate the expectation by drawing random samples from $q(z)$:

$$p(D) \approx \frac{1}{K}\sum_{k=1}^{K} w_k \quad \text{with} \quad z_k \sim q(z)$$

where the quantities

$$w_k = \frac{p(D \mid z_k)p(z_k)}{q(z_k)}$$

are often referred as importance weights and they account for the facts that we are sampling from an arbitrary distribution $q(z)$ instead of the posterior by re-weighting the samples before averaging. This is an example of importance sampling: a general technique to compute integrals using random samples from a fixed distribution.

Exercise 2.1 Consider the importance sampling estimator with K samples:

$$\Phi = \frac{1}{K} \sum_{k=1}^{K} w_k \quad \text{with} \quad z_k \sim q(z)$$

prove that the expectation of this estimator is the marginal likelihood.

Exercise 2.2 Prove that for K tending to infinity the estimator Φ follows an approximate asymptotic normal distribution:

$$\Phi \sim \mathcal{N}\left(p(D), \nu^2/K\right)$$

where ν^2 is the variance of the weight w_k (which we assume to exist).

0.0.3 Evidence lower bound

The problem with importance sampling is that the scale of the weights tend to differ by order of magnitudes so that only a few sample tend to dominate the average. For example, consider a normal posterior distribution

$$p(z \mid D) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(z-\mu_p)^2/2\sigma^2}$$

and a normal sampling distribution with the same variance and different mean:

$$q(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(z-\mu_q)^2/2\sigma^2} \ .$$

in this example, the importance weights are

$$w_k = \frac{p(z_k \mid D)}{q(z_k)} \propto e^{-z_k(\mu_p-\mu_q)/\sigma^2} \ .$$

The exponential function in these weights convert small differences in the values of z_k into enormous differences in the weights. This implies that our importance sampling estimates have a disproportionate dependence on very few samples, leading to very high variance of the estimator.

Exercise 2.3 Find the variance of the weights.

How can we solve this problem? Ideally, we would like to apply a logarithmic function to the weights so to dampen their exponential behavior. We can start by considering the log marginal likelihood:

$$\log P(D) = \log \left(\mathbb{E}_{z \sim q(z)} \left[\frac{p(D \mid z)p(z)}{q(z)} \right] \ . \right) \tag{8}$$

Since the logarithm is a concave function, we can obtain a lower bound of the log marginal likelihood by using Jansen's inequality and thereby move the log inside the expectation:

$$\log p(D) = \log\left(\mathbb{E}_{z\sim q(z)}\left[\frac{p(D \mid z)p(z)}{q(z)}\right]\right) \ , \tag{9}$$

$$\geq \mathbb{E}_{z\sim q(z)}\left[\log\left(\frac{p(D \mid z)p(z)}{q(z)}\right)\right] \ , \tag{10}$$

$$\approx \frac{1}{K}\sum_{k}^{K} \log w_k \ . \tag{11}$$

This quantity is appropriately referred to as Evidence Lower BOund (ELBO). By taking the exponential of both sides, we obtain a lower bound for the marginal likelihood:

$$p(D) \geq e^{\mathbb{E}_{z\sim q(z)}\left[\log\left(\frac{p(D|z)p(z)}{q(z)}\right)\right]} \ , \tag{12}$$

$$\approx e^{\frac{1}{K}\sum_{k}^{K}\log w_k} \ . \tag{13}$$

This quantity is now well-behaved since the log of the weights scale linearly. However, this estimator is now biased and its expectation is only guaranteed to be a lower bound of the real log marginal likelihood. The variance of an importance sampling estimator crucially depends on how close the sampling distribution $q(z)$ is to the real posterior distribution $p(z \mid D)$. In fact, if we use the real posterior as sampling distribution all the weights become deterministically equal to the marginal likelihood:

$$w_k = \frac{p(D \mid z_k)p(z_k)}{p(z_k \mid D)} = \frac{p(D \mid z_k)p(z_k)p(D)}{p(D \mid z_k)p(z_k)} = p(D) \ .$$

Interestingly, this implies that the ELBO is equal to the real log marginal likelihood when we use the posterior as sampling distribution:

$$\frac{1}{K}\sum_{k}^{K}\log w_k = \frac{1}{K}\sum_{k}^{K}\log p(D) = \log p(D) \ . \tag{14}$$

This is a very powerful result, it implies that if we have a "good" sampling distribution $q(z)$ we can estimate the marginal likelihood with the well-behaved estimator in Eq. 0.0.3.

0.0.4 What is a good sampler? The Kullback–Leibler divergence

How can we find a good sampling distribution $q(z)$? Ideally, $q(z)$ should be very similar to the (intractable) posterior $p(z \mid D)$ while being tractable. By tractable, I mean that we need to be able to cheaply sample and evaluate its probability density. We can evaluate the "distance" between a sampling distribution and the posterior using a statistical divergence. A statistical divergence $D(q, p)$ is a positive-valued function of two probability distribution that is zero if and only if the distributions are identical. Divergences are a sort of "measuring rules" in the space of probability distributions. However, they do not need to be symmetrical, meaning that $D(q, p)$ can be different from $D(p, q)$. For our purposes, the most relevant statistical divergence is the Kullback–Leibler divergence (KL-divergence). The KL divergence between two distributions $p_1(z)$ and $p_2(z)$ is defined as follows:

$$D_{KL}\big(p_1(z), p_2(z)\big) = \mathbb{E}_{z\sim p_1(z)}\left[\log \frac{p_1(z)}{p_2(z)}\right] \ . \tag{15}$$

To prove that this function is actually a valid divergence, we need to show that I) it is always non-negative and that II) it is zero only when the distributions are equal. The second property follows from the fact that $\log 1 = 0$. We can prove the first property by noticing that $-\log x \geq 1 - x$ for all values of x, therefore:

$$\begin{aligned} D_{KL}\big(p_1(z), p_2(z)\big) &= \mathbb{E}_{z\sim p_1(z)}\left[-\log \frac{p_2(z)}{p_1(z)}\right] & (16)\\ &\geq \mathbb{E}_{z\sim p_1(z)}\left[1 - \frac{p_2(z)}{p_1(z)}\right] & (17)\\ &= 1 - \int p_1(z)\frac{p_2(z)}{p_1(z)}\mathrm{d}z & (18)\\ &= 1 - 1 = 0 & (19) \end{aligned}$$

Why should we use this specific divergence? It turns out that the KL divergence between sampling distribution and prior is exactly the difference between the ELBO and the true marginal likelihood!

$$\log P(D) = \mathbb{E}_{z\sim q(z)}\left[\log\left(\frac{p(D \mid z)p(z)}{q(z)}\right)\right] + D_{KL}\big(q(z), p(z \mid D)\big) \ . \tag{20}$$

Therefore, the KL divergence quantifies the deviation of our well-behaved estimator (the ELBO) and the true log marginal likelihood.

Exercise 2.4 Prove this result.

0.0.5 Variational inference: Gradient descent in the distribution space

Let's summarize our results: 1) If we have a good sampler we can obtain a well-behaved estimator of the marginal likelihood and 2) we can quantify the quality of a sampler using the KL divergence. What is missing yet is a method to obtain a good sampler $q(z)$. In machine learning and deep learning, we usually solve this kind of problems by defining a parameterized model $q_\theta(z)$ and finding the values of the parameters θ that minimizes a well-chosen loss function.

In this context, the parameterized model is a family of probability distributions whose probability density depends on the parameters. For example, in a univariate problem we could have a normal family parameterized by mean and variance:

$$q_\theta(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(z-\mu)^2/2\sigma^2} \tag{21}$$

where θ is the tuple of parameters (μ, σ^2).

We now need to define a loss function that measures the deviation of our parameterized distribution from the true posterior. In the previous section we learned that the KL divergence is an ideal candidate to fulfil this role:

$$\begin{aligned} \mathcal{L}(\theta) &= D_{KL}\big(q_\theta(z), p(z \mid D)\big) && (22) \\ &= -\mathbb{E}_{z \sim q_\theta(z)}\left[\log \frac{p(z \mid D)}{q_\theta(z)}\right] \ . && (23) \end{aligned}$$

Unfortunately we cannot evaluate this expression since the posterior $P(x \mid D)$ is intractable. If we could, we would already know the marginal likelihood and we would not need any sampler! However, let's for now ignore this problem and try to move forward. We need to evaluate the gradient of the loss function in order to optimize θ by stochastic gradient descent. Let's therefore try to evaluate the gradient:

$$\begin{aligned} \nabla_\theta \mathcal{L}(\theta) &= -\nabla_\theta \mathbb{E}_{z \sim q_\theta(z)}\left[\log \frac{p(z \mid D)}{q_\theta(z)}\right] && (24) \\ &= -\nabla_\theta \mathbb{E}_{z \sim q_\theta(z)}\left[\log \frac{p(z\, D)}{p(D) q_\theta(z)}\right] && (25) \\ &= -\nabla_\theta \left(\mathbb{E}_{z \sim q_\theta(z)}\left[\log \frac{p(z, D)}{q_\theta(z)}\right] + \log P(D)\right) && (26) \\ &= -\nabla_\theta \mathbb{E}_{z \sim q_\theta(z)}\left[\log \frac{p(z, D)}{q_\theta(z)}\right] - \nabla_\theta \log P(D) && (27) \\ &= \nabla_\theta \left(-\mathbb{E}_{z \sim q_\theta(z)}\left[\log \frac{p(z, D)}{q_\theta(z)}\right]\right) \ . && (28) \end{aligned}$$

It turns out that we do not need to know the marguinal likelihood since it only adds a constant shift to the KL divergence which does not contribute to the gradient! In fact, this gradient is identical to the gradient of the ELBO in Eq. 0.0.3. Therefore, we can minimize the negative KL divergence by maximizing the variational log evidence lower bound. For historical reason, this approach is usually referred to as *variational inference* and $q_\theta(z)$ is usually called variational distribution.

Exercise 2.5 Consider the following normal joint distribution

$$\log p(z, D) = -z^2/2 + -\sum_k^K (x_k - z)^2/2 - \frac{(K+1)}{2} \log 2\pi$$

where D is a set of data points $(x_1, ..., x_K)$. We would like to approximate the posterior distribution $p(z \mid D)$ by minimizing the KL between the posterior and a normal sampling distribution as in Eq 0.0.5.

- Compute the gradient of the negative ELBO (Eq. 0.0.6) with respect to the mean and variance parameter. You will need to compute an expected value in closed-form.
- Set those gradients to zero in order to obtain a closed form solution.
- Compare this approximate solution with the true posterior distribution which you should be able to evaluate in closed-form.

0.0.6 Making sense of the ELBO

The variational evidence lower bound can be nicely decomposed into two very interpretable additive terms:

$$\begin{aligned}
\mathcal{L}_{\text{ELBO}}(\theta) &= \mathbb{E}_{z\sim q_\theta(z)}\left[\log \frac{p(z,D)}{q_\theta(z)}\right] && (29)\\
&= \mathbb{E}_{z\sim q_\theta(z)}\left[\log \frac{p(D \mid z)p(z)}{q_\theta(z)}\right] && (30)\\
&= \mathbb{E}_{z\sim q_\theta(z)}[\log p(D \mid z)] - D_{KL}\big(q_\theta(z), p(z)\big) && (31)\\
&= \mathcal{L}_{\text{lk}}(\theta) - \mathcal{R}(\theta) && (32)
\end{aligned}$$

The first term is the average log likelihood under the sampling distribution:

$$\mathcal{L}_{\text{lk}}(\theta) = -\mathbb{E}_{z\sim q_\theta(z)}[\log p(D \mid z)] \ . \tag{33}$$

this is a probabilistic generalization of the standard maximum likelihood losses used in deep learning and in many other fields such as statistics and engineering. Minimizing this loss concentrates all the probability density of the sampler around the values of z that maximize the likelihood. However, using this loss alone would lead to a severe underestimation of the uncertainty in the posterior as all the density would concentrate to its points of maximum (MAP estimate).

The concentration tendency of the loss is counteracted by the regularization term:

$$\mathcal{R}(\theta) = D_{KL}\big(q_\theta(z), p(z)\big) \ . \tag{34}$$

This term forces the sampler to not deviate too much from the prior distribution $p(z)$. This term can be itself be decomposed into $E[\log p(z)]$ which pushes the samples towards the point with highest prior density and $E[-\log q(z)]$ which is the entropy of the sampler and avoids the samples to collapse into a single point-estimate. This can be seen as a repulsion force between the samples, as visualized in Fig. 3.

0.0.7 Stochastic gradient estimation and the reparameterization trick

In order to compute the gradient of the ELBO, we need to evaluate an expectation with respect to sampler:

$$\nabla_\theta \mathcal{L}_{\text{ELBO}}(\theta) = \nabla_\theta \mathbb{E}_{z\sim q_\theta(z)}\left[\log \frac{p(z,D)}{q_\theta(z)}\right] \ . \tag{35}$$

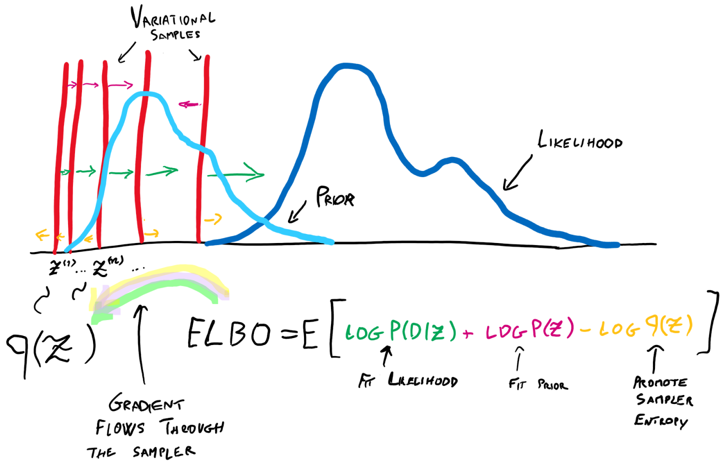

Figure 3: Visualization of the ELBO.

Note that we cannot move the gradient inside the expectation since the sampler itself depends on θ and neglecting this dependency would lead to an incorrect gradient. This is very clear in the case of the probabilistic log likelihood loss:

$$\nabla_\theta \mathcal{L}_{\mathrm{lk}}(\theta) = -\nabla_\theta \mathbb{E}_{z \sim q_\theta(z)}[\log p(D \mid z)] \ . \tag{36}$$

If we push the derivative inside the expectation, we will get:

$$\mathbb{E}_{z \sim q_\theta(z)}[\nabla_\theta \log p(D \mid z)] \ , \tag{37}$$

which is equal to zero since the likelihood does not depend on the variational parameters. The correct gradient should push the sampler towards the values of z associated with high likelihood. Therefore, in order to compute this gradient we need to differentiate (or backpropagate in deep learning lingo) through the sampling itself.

How can we do that? An option is to "reparaterize" our sampler as a deterministic transformation $f_\theta(\epsilon)$ of a fixed sampling distribution $q_0(\epsilon)$. Namely, we need to find a parameterized differentiable function $f_\theta(\epsilon)$ and a distribution $q_0(\epsilon)$ such that:

$$f_\theta(\epsilon) \sim q_\theta(z), \quad \text{for} \quad \epsilon \sim q_0(z) \tag{38}$$

For example, we can reparameterize a normal distribution parameterized by $\theta = (\mu, \sigma)$ as an affine transformation of a standard normal distribution:

$$f_\theta(\epsilon) = \sigma\epsilon + \mu \sim \mathcal{N}(\mu, \sigma^2), \quad \text{for} \quad \epsilon \sim \mathcal{N}(0, 1) \tag{39}$$

Using a reparameterization, we can express the gradient in terms of an expectation with respect to a fixed sampling distribution:

$$\nabla_\theta \mathcal{L}_{\text{ELBO}}(\theta) = \nabla_\theta \mathbb{E}_{z \sim q_\theta(z)} \left[\log \frac{p(z, D)}{q_\theta(z)} \right] \tag{40}$$

$$= \nabla_\theta \mathbb{E}_{\epsilon \sim q_0(\epsilon)} \left[\log \frac{p(f_\theta(\epsilon), D)}{q_\theta(f_\theta(\epsilon))} \right] \ . \tag{41}$$

We can now push the gradient inside the expectation since the sampler does not depend on θ:

$$\nabla_\theta \mathcal{L}_{\text{ELBO}}(\theta) = \mathbb{E}_{\epsilon \sim q_0(\epsilon)} \left[\nabla_\theta \log \frac{p(f_\theta(\epsilon), D)}{q_\theta(f_\theta(\epsilon))} \right] \ . \tag{42}$$

Usually we cannot evaluate the expectation in closed form. However, we can obtain a unbiased Monte Carlo estimate of the gradient by sampling from the base distribution:

$$\nabla_\theta \mathcal{L}_{\text{ELBO}}(\theta) \approx \frac{1}{N} \sum_n^N \nabla_\theta \log \frac{p(f_\theta(\epsilon^{(n)}), D)}{q_\theta(f_\theta(\epsilon^{(n)}))} \quad \text{for} \quad \epsilon^{(n)} \sim q_0(\epsilon) \ . \tag{43}$$

0.0.8 Stochastic convergence

We now have a Monte Carlo estimator of the gradient of the negative ELBO:

$$\mathcal{G}^{(N)}(\theta) = -\frac{1}{N} \sum_n^N \nabla_\theta \log \frac{p(f_\theta(\epsilon^{(n)}), D)}{q_\theta(f_\theta(\epsilon^{(n)}))} \quad \text{for} \quad \epsilon^{(n)} \sim q_0(\epsilon) \tag{44}$$

It is easy to check that this estimator is unbiased, meaning that its average is the true gradient of the ELBO. The variance of the estimator depends both on the number N of samples and on the choice of reparameterization.

We can now use this estimator in a stochastic gradient descent rule:

$$\theta_{n+1} = \theta_n - \eta_n \mathcal{G}^{(M)}(\theta_n) \tag{45}$$

where γ_n is a positive adaptive learning rate. The resulting algorithm is guaranteed to converge to a local minimum when:

$$\sum_{n=1}^{\infty} \eta_n = \infty \tag{46}$$

while

$$\sum_{n=1}^{\infty} \eta_n^2 \leq \infty \ . \tag{47}$$

This suggests the use of a scaling rule of the form:

$$\gamma_n = \frac{\alpha}{1 + \delta n} \ . \tag{48}$$

However, this is only guaranteed to converge to a local minimum. Unfortunately, in variational inference, like in other areas of deep learning, we have no guarantees to to achieve convergence to the best possible approximation. A human inspection of the final result is therefore required.

Furthermore, the learning rate requirements in Eq. 45 and Eq. 45 are only asymptotic and they are not a very useful guidance in practice. In our applications, we will therefore use standard deep learning optimizes such as Adam.

0.0.9 The REINFORCE gradient estimator

It is not always possible to find a reparameterization of the variational sampler. For example, it is clearly impossible to exactly reparameterize discrete distributions since their sampling is not a differentiable process.

In these cases, we can obtain another gradient estimator that does not require reparameterization. We start by writing the expectation as an integral and pushing the gradient inside:

$$\nabla_z \mathbb{E}_{z \sim q_\theta(z)}[g_\theta(z)] = \int \left(\nabla_\theta q_\theta(z) g_\theta(z) + q_\theta(z) \nabla_\theta g_\theta(z) \right) \mathrm{d}z \tag{49}$$

$$= \int \nabla_\theta q_\theta(z) g_\theta(z) \mathrm{d}z + \mathbb{E}_{z \sim q_\theta(z)}[\nabla_\theta g_\theta(z)] \ . \tag{50}$$

where $g_\theta(z) = \log\left(p(z, D)/q_\theta(z)\right)$. Unfortunately, the first term in this expression does not have the form of an expectation integral. This is problematic since our aim is to obtain a Monte Carlo estimator by replacing the exact expectation with an average over finite number of samples. However, we can remedy this this apparent problem by noticing that

$$\nabla_\theta q_\theta(z) = q_\theta(z) \nabla_\theta \log q_\theta(z) \ . \tag{51}$$

We can now plug in this formula and obtain a reparameterization-free expectation formula for the gradient:

$$\nabla_z \mathbb{E}_{z \sim q_\theta(z)}[g_\theta(z)] = \mathbb{E}_{z \sim q_\theta(z)}[\nabla_\theta \left((\log q_\theta(z) + 1) g_\theta(z) \right)] \ . \tag{52}$$

We can finally turn the formula into a Monte Carlo gradient estimator:

$$\mathcal{G}_R^{(N)}(\theta) = -\frac{1}{N} \sum_n^N \nabla_\theta \left((\log q_\theta(z_n) + 1) g_\theta(z_n) \right) \quad \text{for} \quad z^{(n)} \sim q(z) \ . \tag{53}$$

It is easy to see that this estimator is unbiased. However, its variance tend to be much higher than a reparameterization estimator. Therefore, this REINFORCE method should only be used as last resort.

0.0.10 The mean field approximation

How do we chose an appropriate sampling distribution $q_\theta(z)$? Consider a multivariate Bayesian inference problem with a a set of latent variables $z_1, \dots, z_M$

following a multivariate prior distribution $p(z_1, \dots, z_M)$. It is challenging to design a properly parameterized multivariate density for the sampler. A simpler and widely adopted solution is to assume that all the latent variables are statistically independent:

$$q(z_1, \dots, z_M) = \prod_m^M q_m(z_n) \tag{54}$$

We then need to assume an appropriate parameterization for the variational marginals $q_n(z_n)$. For example, if all the latent variables are defined over real numbers we can assume them to be normal distributions parameterized by mean and variance:

$$q(z_1, \dots, z_M) = \prod_m^M \frac{1}{\sqrt{2\pi\nu_m}} e^{-(z_n - \mu_m)^2/2\nu_n} \tag{55}$$

where $\mu_1, ..., \mu_M$ and $\nu_1, ..., \nu_M$ are the learnable parameters of our variational sampler. As we saw in the previous sections, we can train these parameters by minimizing the negative ELBO by stochastic gradient descent using the reparameterization trick:

$$\mathcal{L}_{\text{MF}} = -\frac{1}{N} \sum_{n=1}^N \log p(D, \dots, \mu_m + \sqrt{\nu_m} \epsilon_m^{(n)}, \dots) \tag{56}$$

$$+ \frac{1}{N} \sum_{n=1}^N \sum_m^M \log q_m(\mu_m + \sqrt{\nu_m} \epsilon_m^{(n)}) \tag{57}$$

where each $\epsilon_m^{(n)}$ is sampled independently from a standard normal distribution. Ideally, we would like this approximation to match the true posterior marginal distributions:

$$q_n(z_n) \approx p(z_n \mid D) \ . \tag{58}$$

Unfortunately, the true marginal posteriors do not minimize the negative ELBO for a mean field family if the posterior has correlations between variables. It is very easy to see this when both the true posterior and the mean field variational sampler follow a normal distribution. Consider the following bivariate posterior:

$$p(z_1, z_2 \mid D) \propto e^{-z_1^2/2\sigma^2 - z_2^2/2\sigma^2 + \rho z_1 z_2/\sigma^2} \ , \tag{59}$$

where ρ is the correlation coefficient between z_1 and z_2. Consider now a mean field normal variational distribution with zero mean and learnable variance ν:

$$q(z_1, z_2 \mid D) = \frac{1}{2\pi\nu} e^{-z_1^2/2\nu - z_2^2/2\nu} \ . \tag{60}$$

Up to additive terms that are constant in the learnable parameter ν, the KL divergence between the variational sampler and the posterior is given by the closed form expression:

$$\text{KL}(q, p) \propto -\log \nu^2 + \frac{2\nu}{\sigma^2(1 - \rho^2)} \ . \tag{61}$$

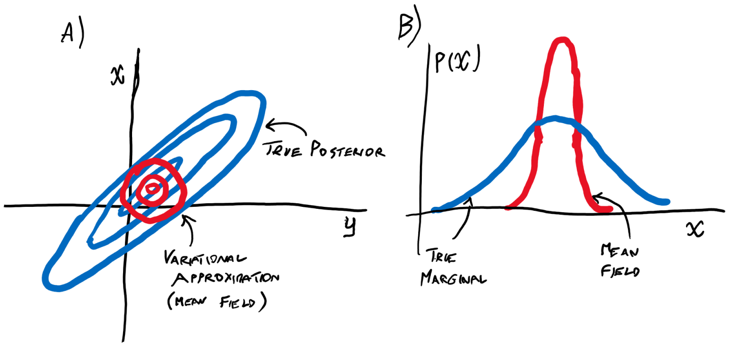

Figure 4: Marginals and the mean field approximations.

We can find the optimal variational parameters by minimizing the KL divergence (note that this is equivalent to minimizing the ELBO):

$$\frac{\partial \mathrm{KL}}{\partial \nu} \propto +\frac{2}{\nu} - \frac{2}{\sigma^2(1-\rho^2)} = 0 \Longrightarrow \nu \; . \tag{62}$$

Therefore, by setting the derivative equal to zero we obtain

$$\nu = \sigma^2(1-\rho^2) \; . \tag{63}$$

As you can see, the optimized variance of the variational distribution differs from the true marginal variance σ^2 by a shrinking factor that depends on the correlation ρ. This can lead to a very severe underestimation of the true uncertainty when using a mean field variational family.