**Auto-Encoding Variational Bayes**

Diederik P. Kingma, Max Welling

**Abstract**

How can we perform efficient inference and learning in directed probabilistic models, in the presence of continuous latent variables with intractable posterior distributions, and large datasets? We introduce a stochastic variational inference and learning algorithm that scales to large datasets and, under some mild differentiability conditions, even works in the intractable case. Our contributions is two-fold. First, we show that a reparameterization of the variational lower bound yields a lower bound estimator that can be straightforwardly optimized using standard stochastic gradient methods. Second, we show that for i.i.d. datasets with continuous latent variables per datapoint, posterior inference can be made especially efficient by fitting an approximate inference model (also called a recognition model) to the intractable posterior using the proposed lower bound estimator. Theoretical advantages are reflected in experimental results.

다루기 힘든 사후 분포와 대규모 데이터 세트가 있는 연속적인 잠재 변수가 있는 상황에서 방향성 확률 모델에서 효율적인 추론과 학습을 어떻게 수행할 수 있습니까? 우리는 대규모 데이터 세트로 확장되고 약간의 미분 조건에서는 다루기 힘든 경우에도 작동하는 확률적 변이 추론 및 학습 알고리즘을 소개합니다. 우리의 기여는 이중적입니다. 첫째, 변동 하한의 재매개변수화(reparameterization)가 표준 확률적 기울기 방법을 사용하여 직접 최적화할 수 있는 하한 추정기를 산출한다는 것을 보여줍니다. 둘째, i.i.d에 대해 보여줍니다. 데이터 포인트당 연속적인 잠재 변수가 있는 데이터 세트에서 사후 추론은 제안된 하한 추정기를 사용하여 난해한 사후에 근사 추론 모델(인식 모델이라고도 함)을 피팅함으로써 특히 효율적으로 만들 수 있습니다. 이론적 이점은 실험 결과에 반영됩니다.

**1.Introduction**

How can we perform efficient approximate inference and learning with directed probabilistic models whose continuous latent variables and/or parameters have intractable posterior distributions? The variational Bayesian (VB) approach involves the optimization of an approximation to the intractable posterior. Unfortunately, the common mean-field approach requires analytical solutions of expectations w.r.t. the approximate posterior, which are also intractable in the general case. We show how a reparameterization of the variational lower bound yields a simple differentiable unbiased estimator of the lower bound; this SGVB (Stochastic Gradient Variational Bayes) estimator can be used for efficient approximate posterior inference in almost any model with continuous latent variables and/or parameters, and is straightforward to optimize using standard stochastic gradient ascent techniques.

연속 잠재 변수 및/또는 매개변수가 다루기 힘든 사후 분포를 갖는 방향성 확률 모델을 사용하여 효율적인 근사 추론 및 학습을 어떻게 수행할 수 있습니까? VB(variational Bayesian) 접근 방식은 난치성 사후에 대한 근사의 최적화를 포함합니다. 불행히도 일반적인 평균 필드 접근 방식에는 기대치에 대한 분석 솔루션이 필요합니다. 일반적인 경우에도 다루기 힘든 대략적인 후방. 우리는 변동 하한의 재매개변수화가 하한의 단순한 미분 가능한 편향되지 않은 추정치를 산출하는 방법을 보여줍니다. 이 SGVB(Stochastic Gradient Variational Bayes) 추정기는 연속 잠재 변수 및/또는 매개변수가 있는 거의 모든 모델에서 효율적인 근사 사후 추론에 사용할 수 있으며 표준 확률적 기울기 상승 기술을 사용하여 최적화하기 쉽습니다.

For the case of an i.i.d. dataset and continuous latent variables per datapoint, we propose the AutoEncoding VB (AEVB) algorithm. In the AEVB algorithm we make inference and learning especially efficient by using the SGVB estimator to optimize a recognition model that allows us to perform very efficient approximate posterior inference using simple ancestral sampling, which in turn allows us to efficiently learn the model parameters, without the need of expensive iterative inference schemes (such as MCMC) per datapoint. The learned approximate posterior inference model can also be used for a host of tasks such as recognition, denoising, representation and visualization purposes. When a neural network is used for the recognition model, we arrive at the variational auto-encoder.

i.i.d.의 경우 데이터 세트 및 데이터 포인트당 연속 잠재 변수를 고려하여 AEVB(AutoEncoding VB) 알고리즘을 제안합니다. AEVB 알고리즘에서 우리는 SGVB 추정기를 사용하여 인식 모델을 최적화함으로써 추론 및 학습을 특히 효율적으로 만듭니다. 인식 모델을 최적화하면 간단한 조상 샘플링을 사용하여 매우 효율적인 근사 사후 추론을 수행할 수 있습니다. 데이터 포인트당 값비싼 반복적 추론 체계(예: MCMC)가 필요합니다. 학습된 근사 사후 추론 모델은 인식, 잡음 제거, 표현 및 시각화 목적과 같은 여러 작업에도 사용할 수 있습니다. 신경망이 인식 모델에 사용되면 Variational Auto-encoder에 도달합니다.

**2.Method**

The strategy in this section can be used to derive a lower bound estimator (a stochastic objective function) for a variety of directed graphical models with continuous latent variables. We will restrict ourselves here to the common case where we have an i.i.d. dataset with latent variables per datapoint, and where we like to perform maximum likelihood (ML) or maximum a posteriori (MAP) inference on the (global) parameters, and variational inference on the latent variables.

이 섹션의 전략은 연속 잠재 변수가 있는 다양한 방향성 그래픽 모델에 대한 하한 추정기(확률적 목적 함수)를 유도하는 데 사용할 수 있습니다. 여기서는 i.i.d.가 있는 일반적인 경우로 제한합니다. 데이터 포인트당 잠재 변수가 있는 데이터 세트이며 (전역) 매개 변수에 대한 최대 가능성(ML) 또는 최대 사후(MAP) 추론을 수행하고 잠재 변수에 대한 변동 추론을 수행하려는 위치입니다.

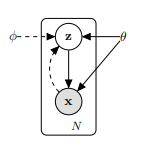


Figure 1: The type of directed graphical model under consideration. Solid lines denote the generative model , dashed lines denote the variational approximation to the intractable posterior . The variational parameters are learned jointly with the generative model parameters θ.

그림 1: 고려 중인 방향성 그래픽 모델의 유형. 실선은 생성 모델 를 나타내고, 점선은 다루기 힘든 사후 에 대한 변동 근사 를 나타냅니다. 변동 매개변수 ϕ는 생성 모델 매개변수 θ와 함께 학습됩니다.

It is, for example, straightforward to extend this scenario to the case where we also perform variational inference on the global parameters; that algorithm is put in the appendix, but experiments with that case are left to future work. Note that our method can be applied to online, non-stationary settings, e.g. streaming data, but here we assume a fixed dataset for simplicity.

예를 들어 이 시나리오를 전역 매개변수에 대한 변이 추론을 수행하는 경우로 확장하는 것은 간단합니다. 그 알고리즘은 부록에 있지만, 그 경우에 대한 실험은 향후 작업으로 남겨둡니다. 우리의 방법은 온라인, 고정되지 않은 설정에 적용될 수 있습니다. 스트리밍 데이터이지만 여기서는 단순성을 위해 고정된 데이터 세트를 가정합니다.

**2.1 Problem scenario**

Let us consider some dataset consisting of N i.i.d. samples of some continuous or discrete variable x. We assume that the data are generated by some random process, involving an unobserved continuous random variable z. The process consists of two steps: (1) a value is generated from some prior distribution a value is generated from some conditional distribution . We assume that the prior and likelihood come from parametric families of distributions and , and that their PDFs are differentiable almost everywhere w.r.t. both θ and z. Unfortunately, a lot of this process is hidden from our view: the true parameters as well as the values of the latent variables are unknown to us.

N i.i.d로 구성된 일부 데이터 세트 을 고려해 보겠습니다. 일부 연속 또는 이산 변수 x의 샘플. 관찰되지 않은 연속 확률 변수 z를 포함하는 임의의 프로세스에 의해 데이터가 생성된다고 가정합니다. 이 프로세스는 두 단계로 구성됩니다. (1) 값 는 일부 사전 분포 에서 생성되고, (2) 값 는 일부 조건부 분포 에서 생성됩니다. 우리는 이전 및 우도 가 및 분포의 매개변수 군에서 유래하고 이들의 PDF가 w.r.t의 거의 모든 곳에서 미분 가능하다고 가정합니다. θ와 z 둘 다. 불행히도, 이 과정의 많은 부분이 우리의 관점에서 숨겨져 있습니다. 진정한 매개변수 과 잠재 변수 의 값은 우리에게 알려지지 않았습니다.

Very importantly, we do not make the common simplifying assumptions about the marginal or posterior probabilities. Conversely, we are here interested in a general algorithm that even works efficiently in the case of:

매우 중요하게, 우리는 한계 또는 사후 확률에 대한 일반적인 단순화 가정을 하지 않습니다. 반대로 다음과 같은 경우에도 효율적으로 작동하는 일반 알고리즘에 관심이 있습니다.

1. Intractability: the case where the integral of the marginal likelihood is intractable (so we cannot evaluate or differentiate the marginal likelihood), where the true posterior density is intractable (so the EM algorithm cannot be used), and where the required integrals for any reasonable mean-field VB algorithm are also intractable. These intractabilities are quite common and appear in cases of moderately complicated likelihood functions , e.g. a neural network with a nonlinear hidden layer.

1. 난치성: 한계 우도 의 적분이 다루기 힘든 경우(따라서 한계 우도를 평가하거나 미분할 수 없음), 여기서 실제 사후 밀도 는 다루기 힘들고(EM 알고리즘을 사용할 수 없음) 합리적인 평균 필드 VB 알고리즘에 필요한 적분도 다루기 어렵습니다. 이러한 난해성은 매우 일반적이며 적당히 복잡한 우도 함수 의 경우에 나타납니다. 비선형 은닉층이 있는 신경망.

2. A large dataset: we have so much data that batch optimization is too costly; we would like to make parameter updates using small minibatches or even single datapoints. Samplingbased solutions, e.g. Monte Carlo EM, would in general be too slow, since it involves a typically expensive sampling loop per datapoint.

2. 대용량 데이터 세트: 데이터가 너무 많아서 일괄 최적화에 비용이 너무 많이 듭니다. 우리는 작은 미니배치 또는 단일 데이터 포인트를 사용하여 매개변수 업데이트를 만들고 싶습니다. 샘플링 기반 솔루션, 예: Monte Carlo EM은 일반적으로 데이터 포인트당 비용이 많이 드는 샘플링 루프를 포함하기 때문에 일반적으로 너무 느립니다.

We are interested in, and propose a solution to, three related problems in the above scenario:

우리는 위의 시나리오에서 세 가지 관련 문제에 관심이 있으며 이에 대한 해결책을 제안합니다.

1. Efficient approximate ML or MAP estimation for the parameters . The parameters can be of interest themselves, e.g. if we are analyzing some natural process. They also allow us to mimic the hidden random process and generate artificial data that resembles the real data.

1. 매개변수 θ에 대한 효율적인 근사 ML 또는 MAP 추정. 매개변수 자체가 관심을 가질 수 있습니다. 우리가 어떤 자연스러운 과정을 분석한다면. 또한 숨겨진 무작위 프로세스를 모방하고 실제 데이터와 유사한 인공 데이터를 생성할 수 있습니다.

2. Efficient approximate posterior inference of the latent variable given an observed value x for a choice of parameters . This is useful for coding or data representation tasks.

2. 매개변수 θ의 선택에 대해 관찰된 값 x가 주어지면 잠재 변수 z의 효율적인 근사 사후 추론. 이것은 코딩 또는 데이터 표현 작업에 유용합니다.

3. Efficient approximate marginal inference of the variable x. This allows us to perform all kinds of inference tasks where a prior over x is required. Common applications in computer vision include image denoising, inpainting and super-resolution.

3. 변수 x의 효율적인 근사 주변 추론. 이를 통해 x보다 우선순위가 필요한 모든 종류의 추론 작업을 수행할 수 있습니다. 컴퓨터 비전의 일반적인 응용 프로그램에는 이미지 노이즈 제거, 인페인팅 및 초고해상도가 포함됩니다.

For the purpose of solving the above problems, let us introduce a recognition model : an approximation to the intractable true posterior . Note that in contrast with the approximate posterior in mean-field variational inference, it is not necessarily factorial and its parameters are not computed from some closed-form expectation. Instead, we’ll introduce a method for learning the recognition model parameters jointly with the generative model parameters .

위의 문제를 해결하기 위해 처리 불가능한 실제 사후 에 대한 근사인 인식 모델 를 도입하겠습니다. 평균 필드 변이 추론의 근사 사후값과 대조적으로, 반드시 계승일 필요는 없으며 매개변수 는 일부 폐쇄형 기대값에서 계산되지 않습니다. 대신에 생성 모델 매개변수 θ와 함께 인식 모델 매개변수 를 학습하는 방법을 소개합니다.

From a coding theory perspective, the unobserved variables have an interpretation as a latent representation or code. In this paper we will therefore also refer to the recognition model as a probabilistic encoder, since given a datapoint x it produces a distribution (e.g. a Gaussian) over the possible values of the code z from which the datapoint x could have been generated. In a similar vein we will refer to as a probabilistic decoder, since given a code z it produces a distribution over the possible corresponding values of x.

코딩 이론의 관점에서 관찰되지 않은 변수 z는 잠재적 표현 또는 코드로 해석됩니다. 따라서 이 백서에서는 인식 모델 를 확률적 인코더로 참조할 것입니다. 생성되었을 수 있습니다. 비슷한 맥락에서 우리는 를 확률적 디코더로 참조할 것입니다. 주어진 코드 z가 x의 가능한 해당 값에 대한 분포를 생성하기 때문입니다.

**2.2 The variational bound**

The marginal likelihood is composed of a sum over the marginal likelihoods of individual datapoints log ,) = , which can each be rewritten as:

한계 가능성은 개별 데이터 포인트의 한계 가능성에 대한 합으로 구성됩니다. ,) = , 각각 다음과 같이 다시 쓸 수 있습니다.



The first RHS term is the KL divergence of the approximate from the true posterior. Since this KL-divergence is non-negative, the second RHS term ) is called the (variational) lower bound on the marginal likelihood of datapoint i, and can be written as:

첫 번째 RHS 항은 실제 사후에서 근사값의 KL 발산입니다. 이 KL-divergence가 음이 아니므로 두 번째 RHS 항 )는 데이터 포인트 i의 한계 우도에 대한 (변동) 하한이라고 하며 다음과 같이 쓸 수 있습니다.



which can also be written as:

다음과 같이 쓸 수도 있습니다.



We want to differentiate and optimize the lower bound w.r.t. both the variational parameters and generative parameters . However, the gradient of the lower bound w.r.t. is a bit problematic. The usual (na¨ıve) Monte Carlo gradient estimator for this type of problem is: where . This gradient estimator exhibits exhibits very high variance (see e.g. [BJP12]) and is impractical for our purposes.

하한 w.r.t를 미분하고 최적화하려고 합니다. 변동 매개변수 및 생성 매개변수 θ 둘 다. 그러나 하한 w.r.t의 기울기는 는 약간 문제가 있습니다. 이 유형의 문제에 대한 일반적인 (순진한) Monte Carlo 기울기 추정기는 다음과 같습니다. 여기서 . 이 기울기 추정기는 매우 높은 분산을 나타내며(예: [BJP12] 참조) 우리의 목적에 비실용적입니다.

**2.3 The SGVB estimator and AEVB algorithm**

In this section we introduce a practical estimator of the lower bound and its derivatives w.r.t. the parameters. We assume an approximate posterior in the form , but please note that the technique can be applied to the case , i.e. where we do not condition on x, as well. The fully variational Bayesian method for inferring a posterior over the parameters is given in the appendix.

이 섹션에서 우리는 하한과 그 도함수 w.r.t의 실용적인 추정기를 소개합니다. 매개변수. 우리는 q 형식의 대략적인 사후값을 가정하지만, 이 기술은 의 경우, 즉 x에 대해 조건을 지정하지 않는 경우에도 적용할 수 있습니다. 매개변수에 대한 사후 추론을 위한 완전 변이 베이지안 방법은 부록에 나와 있습니다.

Under certain mild conditions outlined in section 2.4 for a chosen approximate posterior  we can reparameterize the random variable using a differentiable transformation of an (auxiliary) noise variable :

섹션 2.4에서 약술한 대략적인 후방 에 대해 설명한 특정 온화한 조건에서 (보조) 노이즈 변수 θ의 미분 가능한 변환 를 사용하여 무작위 변수 를 다시 파라미터화할 수 있습니다.



See section 2.4 for general strategies for chosing such an approriate distribution and function . We can now form Monte Carlo estimates of expectations of some function w.r.t. as follows:

적절한 분포 P(ϵ) 및 함수 를 선택하기 위한 일반적인 전략에 대해서는 섹션 2.4를 참조하십시오. 이제 함수 f(z) w.r.t의 기대치에 대한 몬테카를로 추정치를 형성할 수 있습니다. 는 다음과 같습니다.



We apply this technique to the variational lower bound (eq. (2)), yielding our generic Stochastic Gradient Variational Bayes (SGVB) estimator :

이 기술을 변동 하한(eq. (2))에 적용하여 일반적인 SGVB(Stochastic Gradient Variational Bayes) 추정기 :

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명Algorithm 1 Minibatch version of the Auto-Encoding VB (AEVB) algorithm. Either of the two SGVB estimators in section 2.3 can be used. We use settings M = 100 and L = 1 in experiments.

알고리즘 1 AEVB(Auto-Encoding VB) 알고리즘의 Minibatch 버전입니다. 섹션 2.3의 두 SGVB 추정기 중 하나를 사용할 수 있습니다. 실험에서 설정 M = 100 및 L = 1을 사용합니다.

Often, the KL-divergence of eq. (3) can be integrated analytically (see appendix B), such that only the expected reconstruction error requires estimation by sampling. The KL-divergence term can then be interpreted as regularizing , encouraging the approximate posterior to be close to the prior . This yields a second version of the SGVB estimator , corresponding to eq. (3), which typically has less variance than the generic estimator:

종종 eq.의 KL-분산 는 예상 재구성 오류 만 발생하도록 분석적으로 통합할 수 있습니다(부록 B 참조). 그러면 KL-분산 항은 θ를 정규화하는 것으로 해석될 수 있으며, 대략적인 후방은 이전 에 가깝도록 장려합니다. 그러면 일반적으로 일반 추정기보다 분산이 적은 eq. (3)에 해당하는 두 번째 버전의 SGVB 추정기 가 생성됩니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

Given multiple datapoints from a dataset X with N datapoints, we can construct an estimator of the

marginal likelihood lower bound of the full dataset, based on minibatches:

데이터 포인트가 N개인 데이터 집합 X의 여러 데이터 포인트가 주어지면, 우리는 다음 데이터 포인트의 추정기를 구성할 수 있습니다. 전체 데이터 집합의 한계 우도 하한은 미니배치를 기준으로 합니다.

텍스트이(가) 표시된 사진

자동 생성된 설명

where the minibatch is a randomly drawn sample of M datapoints from the full dataset X with N datapoints. In our experiments we found that the number of samples L per datapoint can be set to 1 as long as the minibatch size M was large enough, e.g. M = 100. Derivatives can be taken, and the resulting gradients can be used in conjunction with stochastic optimization methods such as SGD or Adagrad [DHS10]. See algorithm 1 for a basic approach to compute the stochastic gradients.

여기서 미니 배치 은 N 데이터 포인트가 있는 전체 데이터 세트 X에서 무작위로 추출된 M 데이터 포인트 샘플입니다. 우리의 실험에서 우리는 미니 배치 크기 M이 충분히 크면 데이터 포인트당 샘플 L의 수를 1로 설정할 수 있음을 발견했습니다. M = 100. 도함수 을 사용할 수 있으며 결과 기울기는 SGD 또는 Adagrad[DHS10]와 같은 확률론적 최적화 방법과 함께 사용할 수 있습니다. 확률적 기울기를 계산하는 기본 접근 방식은 알고리즘 1을 참조하세요.

A connection with auto-encoders becomes clear when looking at the objective function given at eq. (7). The first term is (the KL divergence of the approximate posterior from the prior) acts as a regularizer, while the second term is a an expected negative reconstruction error. The function is chosen such that it maps a datapoint and a random noise vector to a sample from the approximate posterior for that datapoint: . Subsequently, the sample is then input to function , which equals the probability density (or mass) of datapoint under the generative model, given . This term is a negative reconstruction error in auto-encoder parlance.

eq에 주어진 목적 함수를 보면 자동 인코더와의 연결이 명확해집니다. (7). 첫 번째 항은 (이전의 대략적인 사후값의 KL 발산)이 정규화 역할을 하는 반면 두 번째 항은 예상되는 음의 재구성 오류입니다. 함수 는 데이터 포인트 및 랜덤 노이즈 벡터 를 해당 데이터 포인트에 대한 대략적인 사후의 샘플에 매핑하도록 선택됩니다.. 그 다음, 샘플 는 확률 밀도(또는 생성 모델에서 데이터 포인트 의 질량), 주어진 . 이 용어는 자동 인코더 용어에서 음의 재구성 오류입니다.

**2.4 The reparameterization trick**

In order to solve our problem we invoked an alternative method for generating samples from . The essential parameterization trick is quite simple. Let z be a continuous random variable, and be some conditional distribution. It is then often possible to express the random variable z as a deterministic variable , where is an auxiliary variable with independent marginal , and is some vector-valued function parameterized by .

문제를 해결하기 위해 우리는 에서 샘플을 생성하는 대체 방법을 호출했습니다. 필수 매개 변수화 요령은 매우 간단합니다. z를 연속형 랜덤 변수이고 를 조건부 분포로 합니다. 그런 다음 종종 랜덤 변수 z를 결정론적 변수 로 표현할 수 있습니다. 여기서 θ는 독립적인 주변 를 갖는 보조 변수이고 는 θ에 의해 매개 변수화된 벡터 값 함수입니다.

This reparameterization is useful for our case since it can be used to rewrite an expectation w.r.t such that the Monte Carlo estimate of the expectation is differentiable w.r.t. . A proofis as follows. Given the deterministic mapping we know that . Therefore1,. It follows that a differentiable estimator can be constructed: . In section 2.3 we applied this trick to obtain a differentiable estimator of the variational lower bound.

이러한 재파라미터화는 기대치의 몬테 카를로 추정치가 미분이 되도록 기대치 w.r.t 를 다시 쓰는 데 사용될 수 있기 때문에 우리의 경우에 유용합니다. 증거는 다음과 같습니다. 결정론적 매핑 가 주어졌을 때 . 를 알 수 있습니다. 따라서 입니다. 따라서 미분 가능한 추정기를 구성할 수 있습니다. 。 섹션 2.3에서는 이 방법을 적용하여 변동 하한을 미분할 수 있는 추정치를 구했습니다.

Take, for example, the univariate Gaussian case: . In this case, a valid reparameterization is , where is an auxiliary noise variable . Therefore,.

예를 들어, 일변량 가우스 경우를 예로 들어, 이라고 가정합니다. 이 경우 유효한 재파라미터화는 이며, 여기서 는 보조 노이즈 변수 ~ N(0,1)입니다. 따라서, 입니다. 여기서, 는 이다.

For which can we choose such a differentiable transformation and auxiliary variable

? Three basic approaches are:어떤 에 대해 미분 가능한 변환 및 보조 변수를 선택할 수 있습니까?

? 세 가지 기본 접근 방식은 다음과 같습니다.

1. Tractable inverse CDF. In this case, let , and let be the inverse CDF of . Examples: Exponential, Cauchy, Logistic, Rayleigh, Pareto, Weibull, Reciprocal, Gompertz, Gumbel and Erlang distributions.

1. 다루기 쉬운 역 CDF입니다. 이 경우 로 하고 를 의 역 CDF로 합니다. 예를 들어 다음과 같습니다. 지수 분포, 코치 분포, 로지스틱 분포, 레일리 분포, 파레토 분포, 와이불 분포, 상호 분포, 곰퍼츠 분포, 곰벨 분포 및 얼랑 분포입니다.

2. Analogous to the Gaussian example, for any ”location-scale” family of distributions we can choose the standard distribution (with location = 0, scale = 1) as the auxiliary variable, and let g(.) = location + scale · . Examples: Laplace, Elliptical, Student’s t, Logistic,Uniform, Triangular and Gaussian distributions.

2. 가우스 예시와 유사하게, 모든 "위치 척도" 분포군에 대해 표준 분포(위치 = 0, 척도 = 1)를 보조 변수 로 선택하고 g(.) = 위치 + 척도 · . 예: 라플라스, 타원형, 학생 t, 로지스틱, 균일, 삼각 및 가우스 분포로 설정할 수 있습니다.

3. Composition: It is often possible to express random variables as different transformations of auxiliary variables. Examples: Log-Normal (exponentiation of normally distributed variable), Gamma (a sum over exponentially distributed variables), Dirichlet (weighted sum of Gamma variates), Beta, Chi-Squared, and F distributions.3. 구성은 다음과 같습니다. 종종 랜덤 변수를 보조 변수의 다른 변환으로 표현할 수 있습니다. 예: 로그 정규 분포(정규 분포 변수의 증가), 감마(지수 분포 변수의 합), 디리클레(감마 변수의 가중치 합), 베타, 카이-제곱 및 F 분포입니다.

When all three approaches fail, good approximations to the inverse CDF exist requiring computations with time complexity comparable to the PDF (see e.g. [Dev86] for some methods).

세 가지 접근 방식이 모두 실패하면, PDF에 버금가는 시간 복잡도의 계산이 필요한 역 CDF에 대한 근사치가 존재합니다(예: 일부 방법은 [Dev86] 참조).

**3 Example: Variational Auto-Encoder**

In this section we’ll give an example where we use a neural network for the probabilistic encoder (the approximation to the posterior of the generative model ) and where the parameters and are optimized jointly with the AEVB algorithm.이 섹션에서는 확률론적 인코더 (생성 모델 의 후방에 대한 근사치)에 대해 신경망을 사용하고 매개변수 와 가 AEVB 알고리즘과 함께 최적화되는 예를 제시하겠습니다.

Let the prior over the latent variables be the centered isotropic multivariate Gaussian . Note that in this case, the prior lacks parameters. We let be a multivariate Gaussian (in case of real-valued data) or Bernoulli (in case of binary data) whose distribution parameters are computed from z with a MLP (a fully-connected neural network with a single hidden layer, see appendix C). Note the true posterior is in this case intractable. While there is much freedom in the form , we’ll assume the true (but intractable) posterior takes on a approximate Gaussian form with an approximately diagonal covariance. In this case, we can let the variational approximate posterior be a multivariate Gaussian with a diagonal covariance structure2

그 전은 잠재적인 변수에 대한 중심 등방성 다변량 가우시안자.이 사건에는 이전 변수가 부족합니다.우리는 는 다변수의 가우시안(실수치의 데이터의 경우)또는 베르누이(이진 데이터의 경우)가 상대적으로 매개 변수 z에서 소수 언어 보호(한 숨겨진 계층과fully-connected 신경 조직, 부록 C 보)로 계산되게 하다.진정한 후 이 경우는 다루기 어렵습니다.거기에 있는 동안 형태 에서 많은 자유는, 우리는 진정한( 하지만 난치)posterior 거의 대각선 공분산을 대략적인 가우스 양식을 떠맡는다고 가정할 것이다.이런 경우에, 우리는 변분 개산 posterior 다변량 가우시안 대각선 공분산 structure2과 쉴 수 있습니다.:



where the mean and s.d. of the approximate posterior, and , are outputs of the encoding MLP, i.e. nonlinear functions of datapoint and the variational parameters (see appendix C).

여기서 대략적인 후방, 및 의 평균 및 s.d.는 부호화 MLP, 즉 데이터 포인트 의 비선형 함수 및 변동 매개변수 (의 출력입니다(부록 C 참조).

As explained in section 2.4, we sample from the posterior . With we signify an element-wise product. In this model both (the prior) and are Gaussian; in this case, we can use the estimator of eq. (7) where the KL divergence can be computed and differentiated without estimation (see appendix B). The resulting estimator for this model and datapoint is:

섹션 2.4에서 설명한 대로 여기서 . ⨀는 요소별 곱을 나타냅니다. 이 모델에서 (사전)와 는 모두 가우스입니다. 이 경우 eq의 추정기를 사용할 수 있습니다. (7) 여기서 KL 발산은 추정 없이 계산되고 미분될 수 있습니다(부록 B 참조). 이 모델 및 데이터 포인트 에 대한 결과 추정기는 다음과 같습니다.



As explained above and in appendix C, the decoding term is a Bernoulli or Gaussian MLP, depending on the type of data we are modelling.

위와 부록 C에서 설명한 바와 같이, 디코딩 용어 는 모델링 중인 데이터의 유형에 따라 Bernouli 또는 Gausian MLP입니다.

**4 Related work**

The wake-sleep algorithm [HDFN95] is, to the best of our knowledge, the only other on-line learning method in the literature that is applicable to the same general class of continuous latent variable models. Like our method, the wake-sleep algorithm employs a recognition model that approximates the true posterior. A drawback of the wake-sleep algorithm is that it requires a concurrent optimization of two objective functions, which together do not correspond to optimization of (a bound of) the marginal likelihood. An advantage of wake-sleep is that it also applies to models with discrete latent variables. Wake-Sleep has the same computational complexity as AEVB per datapoint.

Wake-sleep 알고리즘[HDFN95]은 우리가 아는 한 문헌에서 동일한 일반 클래스의 연속 잠복 변수 모델에 적용할 수 있는 유일한 온라인 학습 방법입니다. 우리의 방법과 마찬가지로 wake-sleep 알고리즘은 실제 후방에 가까운 인식 모델을 사용합니다. 웨이크-슬립 알고리즘의 단점은 두 가지 목적 함수의 동시 최적화가 필요하다는 것입니다. 이 둘은 함께 한계 가능성의 최적화(한계)에 해당하지 않습니다. wake-sleep의 장점은 이산 잠재 변수가 있는 모델에도 적용된다는 것입니다. Wake-Sleep은 데이터 포인트당 AEVB와 동일한 계산 복잡성을 갖습니다.

Stochastic variational inference [HBWP13] has recently received increasing interest. Recently, [BJP12] introduced a control variate schemes to reduce the high variance of the na¨ıve gradient estimator discussed in section 2.1, and applied to exponential family approximations of the posterior. In [RGB13] some general methods, i.e. a control variate scheme, were introduced for reducing the variance of the original gradient estimator. In [SK13], a similar reparameterization as in this paper was used in an efficient version of a stochastic variational inference algorithm for learning the natural parameters of exponential-family approximating distributions.

확률적 변이 추론[HBWP13]은 최근에 증가하는 관심을 받고 있습니다. 최근에 [BJP12]는 섹션 2.1에서 논의된 기본 기울기 추정기의 고분산을 줄이기 위해 제어 변수 계획을 도입했으며 사후의 지수 패밀리 근사에 적용했습니다. [RGB13]에서는 원래 기울기 추정기의 분산을 줄이기 위해 일부 일반적인 방법, 즉 제어 변수 체계가 도입되었습니다. [SK13]에서는 이 논문에서와 유사한 재매개변수화가 지수-패밀리 근사 분포의 자연 매개변수를 학습하기 위한 확률적 변이 추론 알고리즘의 효율적인 버전에 사용되었습니다.

The AEVB algorithm exposes a connection between directed probabilistic models (trained with a variational objective) and auto-encoders. A connection between linear auto-encoders and a certain class of generative linear-Gaussian models has long been known. In [Row98] it was shown that PCA corresponds to the maximum-likelihood (ML) solution of a special case of the linear-Gaussian model with a prior and a conditional distribution, specifically the case with infinitesimally small .

AEVB 알고리즘은 방향성 확률 모델(변동 목표로 훈련됨)과 자동 인코더 간의 연결을 노출합니다. AEVB 알고리즘은 (변동 목표를 가지고 훈련됨) 유도 확률론적 모델과 자동 인코더 사이의 연결을 노출합니다. 선형 자동 인코더와 특정 클래스의 생성 선형-가우스 모델 간의 연결은 오랫동안 알려져 왔습니다. [Row98]에서 PCA는 이전 및 조건부 분포 , 특히 극소수의 경우 선형-가우스 모델의 특수 케이스의 최대우도(ML) 솔루션에 해당함을 보여주었습니다.

In relevant recent work on autoencoders [10] it was shown that the training criterion of unregularized autoencoders corresponds to maximization of a lower bound (see the infomax principle [Lin89]) of the mutual information between input X and latent representation Z. Maximizing (w.r.t. parameters) of the mutual information is equivalent to maximizing the conditional entropy, which is lower bounded by the expected loglikelihood of the data under the autoencoding model [10], i.e. the negative reconstrution error. However, it is well known that this reconstruction criterion is in itself not sufficient for learning useful representations [BCV13]. Regularization techniques have been proposed to make autoencoders learn useful representations, such as denoising, contractive and sparse autoencoder variants [BCV13]. The SGVB objective contains a regularization term dictated by the variational bound (e.g. eq. (10)), lacking the usual nuisance regularization hyperparameter required to learn useful representations. Related are also encoder-decoder architectures such as the predictive sparse decomposition (PSD) [KRL08], from which we drew some inspiration. Also relevant are the recently introduced Generative Stochastic Networks [BTL13] where noisy auto-encoders learn the transition operator of a Markov chain that samples from the data distribution. In [SL10] a recognition model was employed for efficient learning with Deep Boltzmann Machines. These methods are targeted at either unnormalized models (i.e. undirected models like Boltzmann machines) or limited to sparse coding models, in contrast to our proposed algorithm for learning a general class of directed probabilistic models.

자동 인코더에 대한 관련 최근 연구에서, 비정규화된 자동 인코더의 훈련 기준은 입력 X와 잠재 표현 Z 사이의 상호 정보의 하한 (인포맥스 원칙 [Lin89] 참조)의 최대화에 해당하는 것으로 나타났습니다. 상호 정보의 최대화(w.r.t. 매개 변수)는 조건부 엔트로피의 최대화와 같으며, 이는 자동 인코딩 모델 [vVLL+^+10], 즉 음의 재구성 오류에서 데이터의 예상 로그 우도에 의해 경계가 낮아집니다. 그러나, 이 재구성 기준 자체가 유용한 표현을 학습하기에 충분하지 않다는 것은 잘 알려져 있습니다 [BCV13]. 자동 인코더가 노이즈 제거, 수축 및 스파스 자동 인코더 변종과 같은 유용한 표현을 학습하도록 하기 위한 정규화 기법이 제안되었습니다 [BCV13]. SGVB 목표에는 유용한 표현을 학습하는 데 필요한 일반적인 성가신 정규화 하이퍼 매개 변수가 없는 변동 한계(예: (10))에 의해 지시된 정규화 항이 포함되어 있습니다. 이와 관련된 것은 또한 예측 스파스 분해 (PSD) [KRL08]와 같은 인코더-디코더 아키텍처이며, 여기에서 영감을 얻었습니다. 또한 최근에 도입된 생성 확률 네트워크 (BTL13)도 관련이 있습니다. 여기서 노이즈가 많은 자동 인코더는 데이터 분포에서 표본을 추출하는 마르코프 체인의 전환 연산자를 학습합니다. [SL10]에서는 Deep Boltzmann Machines의 효율적인 학습을 위해 인식 모델을 사용했습니다. 이러한 방법은 일반 등급의 유도 확률론적 모델을 학습하기 위해 제안된 알고리즘과 달리 정규화되지 않은 모델(즉, 볼츠만 기계와 같은 무방향 모델) 또는 희박한 코딩 모델로 제한됩니다.

The recently proposed DARN method [GMW13], also learns a directed probabilistic model using an auto-encoding structure, however their method applies to binary latent variables. Even more recently, [RMW14] also make the connection between auto-encoders, directed proabilistic models and stochastic variational inference using the reparameterization trick we describe in this paper. Their work was developed independently of ours and provides an additional perspective on AEVB.최근 제안된 DARN 방법[GMW13]도 자동 인코딩 구조를 사용하여 방향성 확률 모델을 학습하지만 이 방법은 이진 잠재 변수에 적용됩니다. 더 최근에 [RMW14]는 또한 우리가 이 백서에서 설명하는 재매개변수화 트릭을 사용하여 자동 인코더, 방향성 확률 모델 및 확률적 변동 추론 사이를 연결합니다. 그들의 작업은 우리와 독립적으로 개발되었으며 AEVB에 대한 추가 관점을 제공합니다.

**5 Experiments**

We trained generative models of images from the MNIST and Frey Face datasets3 and compared learning algorithms in terms of the variational lower bound, and the estimated marginal likelihood. The generative model (encoder) and variational approximation (decoder) from section 3 were used, where the described encoder and decoder have an equal number of hidden units. Since the Frey Face data are continuous, we used a decoder with Gaussian outputs, identical to the encoder, except that the means were constrained to the interval (0, 1) using a sigmoidal activation function at the decoder output. Note that with hidden units we refer to the hidden layer of the neural networks of the encoder and decoder.

우리는 MNIST 및 Frey Face 데이터 세트3에서 이미지의 생성 모델을 훈련하고 변동 하한 및 추정된 한계 가능성 측면에서 학습 알고리즘을 비교했습니다. 섹션 3의 생성 모델(인코더) 및 변형 근사(디코더)가 사용되었으며, 여기서 설명된 인코더와 디코더는 동일한 수의 은닉 유닛을 가집니다. Frey Face 데이터가 연속적이기 때문에 디코더 출력에서 ​​S자형 활성화 함수를 사용하여 평균이 구간(0, 1)으로 제한된다는 점을 제외하고는 인코더와 동일한 가우스 출력이 있는 디코더를 사용했습니다. 은닉 유닛은 인코더와 디코더 신경망의 은닉 레이어를 참조합니다.

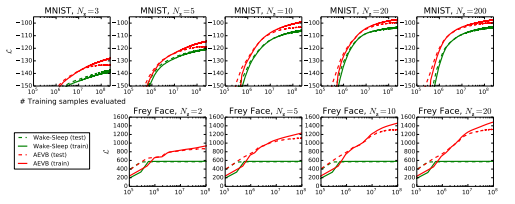


Figure 2: Comparison of our AEVB method to the wake-sleep algorithm, in terms of optimizing the lower bound, for different dimensionality of latent space (). Our method converged considerably faster and reached a better solution in all experiments. Interestingly enough, more latent variables does not result in more overfitting, which is explained by the regularizing effect of the lower bound. Vertical axis: the estimated average variational lower bound per datapoint. The estimator variance was small (< 1) and omitted. Horizontal axis: amount of training points evaluated. Computation took around 20-40 minutes per million training samples with a Intel Xeon CPU running at an effective 40 GFLOPS.

그림 2: 잠복 공간(N\_z)의 다양한 차원에 대한 하한 최적화 측면에서 AEVB 방법과 wake-sleep 알고리즘의 비교. 우리의 방법은 상당히 빠르게 수렴되었고 모든 실험에서 더 나은 솔루션에 도달했습니다. 흥미롭게도, 더 많은 잠재 변수가 더 많은 과적합을 초래하지 않으며, 이는 하한의 정규화 효과로 설명됩니다. 세로 축: 데이터 포인트당 추정된 평균 변동 하한. 추정량 분산이 작고(< 1) 생략되었습니다. 수평 축: 평가된 훈련 포인트의 양. 효과적인 40GFLOPS에서 실행되는 Intel Xeon CPU를 사용하여 백만 개의 훈련 샘플당 계산에 약 20-40분이 소요되었습니다.

Parameters are updated using stochastic gradient ascent where gradients are computed by differentiating the lower bound estimator (see algorithm 1), plus a small weight decay term corresponding to a prior . Optimization of this objective is equivalent to approximate MAP estimation, where the likelihood gradient is approximated by the gradient of the lower bound.

매개변수는 확률적 경사 상승(stochastic gradient accept)을 사용하여 업데이트됩니다. 여기서 구배는 하한 추정기 (알고리즘 1 참조)과 이전 에 해당하는 작은 체중 감소 기간을 더하여 계산됩니다. 이 목표의 최적화는 우도 기울기가 하한의 기울기에 의해 근사되는 대략적인 MAP 추정과 동일합니다.

We compared performance of AEVB to the wake-sleep algorithm [HDFN95]. We employed the same encoder (also called recognition model) for the wake-sleep algorithm and the variational autoencoder. All parameters, both variational and generative, were initialized by random sampling from N (0, 0.01), and were jointly stochastically optimized using the MAP criterion. Stepsizes were adapted with Adagrad [DHS10]; the Adagrad global stepsize parameters were chosen from {0.01,0.02, 0.1} based on performance on the training set in the first few iterations. Minibatches of size M = 100 were used, with L = 1 samples per datapoint.

AEVB의 성능을 웨이크 슬립 알고리즘과 비교하였습니다 [HDFN95]. 웨이크 슬립 알고리즘과 가변 자동 인코더에 동일한 인코더(인식 모델이라고도 함)를 사용했습니다. 모든 변수(변동 및 생성)는 N(0, 0.01)의 무작위 샘플링에 의해 초기화되었으며, MAP 기준을 사용하여 공동으로 확률적으로 최적화되었습니다. 단계 크기는 Adagrad [DHS10]를 사용하여 조정되었습니다. Adagrad 글로벌 단계 크기 매개변수는 처음 몇 번의 반복에서 설정된 교육 결과에 따라 {0.01,0.02,0.1} 중에서 선택되었습니다. 크기가 M = 100인 미니 시계가 사용되었으며, 데이터 포인트당 L = 1개의 샘플이 사용되었습니다.

Likelihood lower bound We trained generative models (decoders) and corresponding encoders (a.k.a. recognition models) having 500 hidden units in case of MNIST, and 200 hidden units in case of the Frey Face dataset (to prevent overfitting, since it is a considerably smaller dataset). The chosen number of hidden units is based on prior literature on auto-encoders, and the relative performance of different algorithms was not very sensitive to these choices. Figure 2 shows the results when comparing the lower bounds. Interestingly, superfluous latent variables did not result in overfitting, which is explained by the regularizing nature of the variational bound.우도 하한 MNIST의 경우 500개의 숨겨진 단위, Frey Face 데이터 집합의 경우 200개의 숨겨진 단위를 가진 생성 모델(디코더) 및 해당 인코더(인식 모델)를 교육했습니다(과적합이 상당히 작기 때문에). 숨겨진 장치의 선택된 수는 자동 인코더의 이전 문헌에 기초하고 있으며, 다른 알고리즘의 상대적 성능은 이러한 선택에 그다지 민감하지 않았습니다. 그림 2는 하한을 비교할 때의 결과를 보여 줍니다. 흥미롭게도 불필요한 잠재 변수는 과적합으로 이어지지 않았으며, 이는 변동 경계의 정규화 특성으로 설명됩니다.

Marginal likelihood For very low-dimensional latent space it is possible to estimate the marginal likelihood of the learned generative models using an MCMC estimator. More information about the marginal likelihood estimator is available in the appendix. For the encoder and decoder we again used neural networks, this time with 100 hidden units, and 3 latent variables; for higher dimensional latent space the estimates became unreliable. Again, the MNIST dataset was used. The AEVB and Wake-Sleep methods were compared to Monte Carlo EM (MCEM) with a Hybrid Monte Carlo (HMC) [DKPR87] sampler; details are in the appendix. We compared the convergence speed for the three algorithms, for a small and large training set size. Results are in figure 3.

한계 가능성 매우 낮은 차원의 잠재 공간에 대해 MCMC 추정기를 사용하여 학습된 생성 모델의 한계 가능성을 추정하는 것이 가능합니다. 주변 우도 추정량에 대한 자세한 내용은 부록에서 확인할 수 있습니다. 인코더와 디코더를 위해 이번에는 100개의 은닉 유닛과 3개의 잠재 변수가 있는 신경망을 다시 사용했습니다. 더 높은 차원의 잠재 공간에 대한 추정치는 신뢰할 수 없게 되었습니다. 다시 MNIST 데이터 세트를 사용했습니다. AEVB 및 Wake-Sleep 방법은 HMC(Hybrid Monte Carlo) [DKPR87] 샘플러를 사용하여 Monte Carlo EM(MCEM)과 비교되었습니다. 자세한 내용은 부록에 있습니다. 우리는 크고 작은 훈련 세트 크기에 대해 세 가지 알고리즘의 수렴 속도를 비교했습니다. 결과는 그림 3에 나와 있습니다.

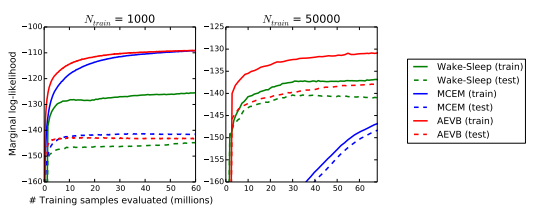


Figure 3: Comparison of AEVB to the wake-sleep algorithm and Monte Carlo EM, in terms of the estimated marginal likelihood, for a different number of training points. Monte Carlo EM is not an on-line algorithm, and (unlike AEVB and the wake-sleep method) can’t be applied efficiently for the full MNIST dataset.

그림 3: 다른 수의 훈련 포인트에 대한 예상 한계 가능성 측면에서 AEVB와 wake-sleep 알고리즘 및 Monte Carlo EM의 비교. Monte Carlo EM은 온라인 알고리즘이 아니며 (AEVB 및 wake-sleep 방법과 달리) 전체 MNIST 데이터 세트에 효율적으로 적용할 수 없습니다.

Visualisation of high-dimensional data If we choose a low-dimensional latent space (e.g. 2D),we can use the learned encoders (recognition model) to project high-dimensional data to a lowdimensional manifold. See appendix A for visualisations of the 2D latent manifolds for the MNIST and Frey Face datasets.

고차원 데이터의 시각화 저차원 잠재 공간(예: 2D)을 선택하면 학습된 인코더(인식 모델)를 사용하여 고차원 데이터를 저차원 매니폴드에 투영할 수 있습니다. MNIST 및 Frey Face 데이터 세트에 대한 2D 잠재 매니폴드의 시각화는 부록 A를 참조하십시오.

**6 Conclusion**

We have introduced a novel estimator of the variational lower bound, Stochastic Gradient VB (SGVB), for efficient approximate inference with continuous latent variables. The proposed estimator can be straightforwardly differentiated and optimized using standard stochastic gradient methods. For the case of i.i.d. datasets and continuous latent variables per datapoint we introduce an efficient algorithm for efficient inference and learning, Auto-Encoding VB (AEVB), that learns an approximate inference model using the SGVB estimator. The theoretical advantages are reflected in experimental results.연속 잠복 변수에 대한 효율적인 근사 추론을 위해 변동 하한 추정기인 확률적 경사 VB(SGVB)를 도입했습니다. 제안된 추정기는 표준 확률적 경사 방법을 사용하여 쉽게 구별하고 최적화할 수 있습니다. i.i.d. 데이터셋과 데이터 포인트당 연속 잠복 변수의 경우 효율적인 추론 및 학습을 위한 효율적인 알고리즘인 자동 인코딩 VB(AEVB)를 도입하여 SGVB 추정기를 사용하여 대략적인 추론 모델을 학습합니다. 이론적 이점은 실험 결과에 반영됩니다.

**7 Future work**

Since the SGVB estimator and the AEVB algorithm can be applied to almost any inference and learning problem with continuous latent variables, there are plenty of future directions: (i) learning hierarchical generative architectures with deep neural networks (e.g. convolutional networks) used for the encoders and decoders, trained jointly with AEVB; (ii) time-series models (i.e. dynamic Bayesian networks); (iii) application of SGVB to the global parameters; (iv) supervised models with latent variables, useful for learning complicated noise distributions.SGVB 추정기 및 AEVB 알고리즘은 연속 잠복 변수의 거의 모든 추론 및 학습 문제에 적용될 수 있기 때문에, (i) 인코더 및 디코더에 사용되는 심층 신경 네트워크(예: 컨볼루션 네트워크)를 사용하여 계층적 생성 아키텍처를 학습하고, EVB와 함께 교육할 수 있습니다. (ii) 시계열 모델 (즉, 동적 베이지안 네트워크), (ii) 글로벌 매개변수에 대한 SGVB 적용, (iv) 복잡한 소음 분포를 학습하는 데 유용한 잠재 변수를 포함하는 감독 모델입니다.