Overview
• We introduce a simple, efficient, and general method for training directed latent variable models.
  – Can handle both discrete and continuous latent variables.
  – Easy to apply – requires no model-specific derivations.
• Key idea: Train an auxiliary neural network to perform inference in the model of interest by optimizing the variational bound.
  – Was considered before for Helmholtz machines and rejected as infeasible due to high variance of inference net gradient estimates.
• We make the approach practical using simple and general variance reduction techniques.
• Promising document modelling results using sigmoid belief networks.

Variational inference
• Given a directed latent variable model that naturally factorizes as
  \[ P(x, h) = P(x|b)P(h), \]
  we can lower-bound the contribution of \( x \) to the log-likelihood as
  \[ \log P(x) \geq \mathbb{E}_Q \left[ \log P(x|b) - \log Q(b|h) \right] = \mathcal{L}_{	heta}(x), \]
  where \( Q(b|h) \) is an arbitrary distribution.
• Variational learning involves alternating between maximizing the lower bound \( \mathcal{L}_{	heta}(x) \) w.r.t. the variational distribution/posterior \( Q(b|h) \) and model parameters \( \theta \).
• Typically variational inference requires:
  – Variational distributions \( Q \) with simple factored form and different parameters for each \( x \).
  – Simple models \( P(x|b) \), yielding tractable expectations.
  – Iterative optimization to compute \( Q \) for each \( x \).

Neural variational inference and learning (NVIL)
• We propose a method that avoids iterative inference, while allowing expressive, potentially multimodal, posteriors and highly expressive models.
• This is achieved by using a feed-forward model for \( Q(b|h) \), making the dependence of the approximate posterior on the input \( x \) parametric.
  – This allows us to sample from \( Q(b|h) \) very efficiently.
  – We refer to \( Q \) as the inference network because it implements approximate inference for the model being trained.
• We train the model and the inference network jointly by updating their parameters to increase the variational lower bound \( \mathcal{L}_{	heta}(x) \).
  – We compute all the required expectations using samples from \( Q \).

Gradients of the variational bound
• The gradients w.r.t. the model and inference net parameters are:
  \[ \frac{\partial}{\partial \theta} \mathcal{L}_{	heta}(x) = \mathbb{E}_Q \left[ \frac{\partial}{\partial \theta} \log P(x,h) \right], \]
  \[ \frac{\partial}{\partial \phi} \mathcal{L}_{\phi}(x) = \mathbb{E}_Q \left[ \log P(x,h) - \log Q(x|h) \right] \frac{\partial}{\partial \phi} \log Q(x|h). \]
• Both gradients can be estimated using samples from the inference net.
• However, the most natural estimator of the inference net gradient is too high-variance to be useful.

Reducing gradient variance
• Key observation: If \( h \) is sampled from \( Q(h|x) \),
  \[ \log P(x,h) - \log Q(h|x) - \mathbb{E}_b \frac{\partial}{\partial \phi} \log Q(h|x), \]
  is an unbiased estimator of \( \frac{\partial}{\partial \phi} \mathcal{L}(x) \) for any \( b \) that does not depend on \( h \).
• Since the variance of the estimator does depend on \( h \), we can obtain estimators with lower variance by choosing \( h \) carefully.
• Our strategy is to choose \( h \) so that the resulting learning signal \( \log P(x,h) - \log Q(h|x) - b \) is close to zero.
• Borrowing terminology from reinforcement learning, we call \( h \) a baseline.

Variance reduction techniques
1. Constant baseline \( b(x) \)
   • Make \( b \) a running estimate of the mean of \( \log P(x,h) - \log Q(h|x) \).
   • Centers the learning signal, making it approximately zero-mean.
   • Enough to obtain reasonable models on MNIST.
2. Input-dependent baseline \( b(x) \)
   • An MLP with a single real-valued output.
   • Can be seen as capturing \( \log P(x,h) \).
   • Makes learning considerably faster and leads to better results.
3. Variance normalization
   • Scale the learning signal to have unit variance.
   • Can be seen as simple global learning rate adaptation.
   • Makes learning faster and more robust.
4. Local learning signals
   • Simpler, less noisy local learning signals can be derived by taking advantage of the Markov properties of the model and the inference net.
   • Likely to be important for training deeper models.

Generative modelling of binarized MNIST
Effect of gradient variance reduction
Figure 1: Sigmoid belief network with 1 hidden layer of 200 units.
Figure 2: Sigmoid belief network with 2 hidden layers of 200 units.

NVIL vs. Wake-Sleep
• SBN is a sigmoid belief network.
• fDARN is an SBN with hidden autoregressive connections.
• Dim is the number of latent variables in each layer, starting with the deepest one.
• NVIL and WS refer to NVIL and wake-sleep training respectively.
• NLL is the negative log-likelihood for the tractable models and an estimate of or a bound on it for the intractable ones.

Document modelling results
• Task: model the joint distribution of word counts in bags of words describing documents.
• Models: SBN and fDARN models with one hidden layer.
• Datasets:
  – 20 Newsroups
  – 11K docs, 2K vocabulary
  – Reuters RCV1
  – 800K docs, 10K vocabulary
• Performance metric: perplexity