Motivation
Trade-off: approximation ability, computational efficiency, and sample complexity.

Contribution:
We proposed coupled variational Bayes (CVB) algorithm that
- hinges the primal-dual view of the ELBO for efficient computation;
- exploits the optimization embedding technique for flexible posterior;
- couples the variational families with the model parameters for better sample complexity.

Preliminary

Variational Bayes: Given a generative model, \( p(x, z) = p_z(z)p_x(x|z) \) with \( x \in \mathbb{R}^d \) as the observations, \( z \in \mathbb{R}^r \) latent variables, and \( \theta \) as the parameters,

\[
\log p_x(x) = \log \int p_z(z)p_x(x|z) \, dz \geq \mathbb{E}_{q(z|x)} \log p_x(x) - \log q_z(z|x).
\]

Existing variational distributions:
- Reparametrized density: Deep model parameterized simple distributions, e.g.,
  \[
  q_z(z|x) = \mathcal{N}(z|\mu_z(x), \text{diag}(\sigma^2_z(x)))
  \]
  with \( \mu_z(x) \) and \( \sigma_z(x) \) as neural networks.
- Disadvantages: Simple densities restrict the approximation ability.
- Tractable flow-based model:
  \[
  q_k(z|x) = q_{z,k}(z|x) \prod_{t=1}^{k} \det \nabla z \theta_t^{-1}
  \]
  with \( z^T = \theta_1 \circ \theta_2 \circ \ldots \circ \theta_k (z) \) and \( z_0 \sim q_0(z|x) \).
- Disadvantages: \( \theta_k \} \) must be invertible; \( \{\theta\} \) \( \det \nabla z \theta_t^{-1} \) is expensive for general \( \theta_t \).

Key Component I: Primal-Dual view of ELBO

Theorem [Primal-Dual reformulation of ELBO] We can reformulate the ELBO as

\[
\max \log p_x(x) = \max \mathbb{E}_{q(z|x)} \log p_x(x) - \log q_z(z|x).
\]

Equivalently as

\[
\min _{\psi_z} \mathbb{E}_{q(z|x)} \left[ \log p(z|x, \psi_z) - \log \psi_z(z) \right] + \mathbb{E}_{p(x, z)} \left[ -\log p(x) \right]
\]

where \( \psi_z \) is \( \{\theta: \mathbb{R}^d \rightarrow \mathbb{R}^r\} \), \( p_z(z) \) denotes some simple distribution and the optimal \( \psi_z^\star (x, z) = \frac{\partial p(x, z)}{\partial q_z(z|x)} \).

Proof sketch: The primal-dual formulation of ELBO is derived based on Fenchel-duality of KL-divergence, i.e.,

\[
KL(q||p) = \max \log q - \min \{p, q\} = \max \{q, \log \nu\} + 1,
\]

and interchangeability principle [Dai et al. 2016].

Nonparametric representation of \( q \): We are able to represent the distributional operation on \( q \) by local variables \( z_{k-1} \).

Efficient computation: We can avoid the computation of \( \log \text{det} \) of Jacobian.

Key Component II: Optimization Embedding

By the Theorem, we can obtain the implicit nonparametric transformation from \( (x, z) \in \mathbb{R}^d \times \mathbb{R}^r \) to \( z_{k-1} \in \mathbb{R}^r \) as

\[
\begin{align*}
  z_{k-1} & = \arg\max_{z_{k-1}} \log p_z(z_{k-1}) - \log q_z(z_{k-1} | x, \psi_z^\star) \\
  & = \text{argmax} \left[ \mathbb{E}_{q_z} \log q_z(z_{k-1} | x, \psi_z^\star) \right]
  \end{align*}
\]

Optimization Embedding with Mirror Descent Algorithm (MDA):

\[
\begin{align*}
  z_{k-1} & = \arg\max_{z_{k-1}} \mathbb{E}_{q_z} \log q_z(z_{k-1} | x, \psi_z^\star) - D_z(z_{k-1}, z) \\
  & = \text{argmax} \left[ \mathbb{E}_{q_z} \log q_z(z_{k-1} | x, \psi_z^\star) + f(z_{k-1}, z) \right]
  \end{align*}
\]

where \( g(z, z_{k-1}^*) = \nabla_z \log p_z(z_{k-1}^*) - \nabla_z \log q_z(z_{k-1}^* | x, \psi_z^\star) \), \( D_z(z, z_{k-1}^*) = \omega(z) - \omega(z_{k-1}^*) - \nabla \omega(z_{k-1}^*) \cdot (z - z_{k-1}^*) \). \( f(z, z_{k-1}^*) = \omega(z) \). The MDA algorithm naturally establishes nonparametric function that maps from \( \mathbb{R}^d \times \mathbb{R}^r \) to \( \mathbb{R}^r \) to approximate the mapping defined by \( T \).

\[
\begin{align*}
  z_{k-1} & \approx z_{k-1}^* = \arg\max_{z_{k-1}} \log p_z(z_{k-1} | x) - \log q_z(z_{k-1} | x, \psi_z^\star).
  \end{align*}
\]

Unbiased Gradient Estimator: We have the surrogate as

\[
\max \log p_x(x) = \max \mathbb{E}_{q(z|x)} \log p_x(x) - \log q_z(z|x).
\]

with \( f(z, \theta) = \mathbb{E}_{q_z} \left[ \log p_z(z | x, \psi_z^\star) \right] - \mathbb{E}_{q_z} \left[ \log q_z(z | x, \psi_z^\star) \right] \). Denote \( \psi^\star (x, z) = \arg\min_{\psi_z} \mathbb{E}_{q_z} f(z, \psi_z) \), we have the unbiased gradient estimator w.r.t. \( \theta \) as

\[
\frac{\partial f(z, \theta)}{\partial \theta} = \mathbb{E}_{q_z} \left[ \nabla_{\theta} \log p_x(x) \right] - \mathbb{E}_{q_z} \left[ \nabla_{\theta} \log q_z(z | x, \psi^\star) \right]
\]

Generalization:
The Optimization Embedding technique can be generalized with arbitrary differentiable algorithms besides the MDA.

Optimization Embedding as Gradient Flow

Theorem: For a continuous time \( t = \eta T \) and infinitesimal step size \( \eta \rightarrow 0 \), the density of the particles \( z \in \mathbb{R}^r \) denoted as \( q_t(z|x) \), follows nonlinear Fokker-Planch equation

\[
\frac{\partial q_t(z|x)}{\partial t} = - \nabla_z \log q_t(z|x) \psi_t(z|x).
\]

such process is a gradient flow of KL-divergence in the space of measures with \( 2 \) Wasserstein metric.

Flexibility in Posterior Approximation

Figure: Distribution of the latent variables for VAE and CVB on synthetic dataset.

Algorithm

Coupled Variational Bayes (CVB)

1. Initialize \( \theta, V \) and \( W \) (the parameters of \( p \) and \( z^2 \)) randomly, set length of steps \( T \) and mirror function \( f \) 
Set \( z(0) = \eta(0) \).
2. for iteration \( k = 1, \ldots, K \) do
3. Sample mini-batch \( \{x_i\}_{i=1}^m \) from dataset \( D \), \( \{z_i\}_{i=1}^m \) from prior \( p(z) \), and \( \{\psi_i\}_{i=1}^m \) from \( p(\theta) \).
4. for iteration \( t = 1, \ldots, T \) do
5. Compute \( z_{t}^t | \psi_t, z \) for each pair of \( (x, z_t) \).
6. - Descend \( V \) with \( \nabla V \) \( \sum_{i=1}^m \left[ \nabla V(x_i, z_i) - \log \nu V(x_i, z_t(z_i)) \right] \).
7. end for
8. Ascend \( \theta \) by stochastic gradient (2).
9. Ascend \( W \) by \( \nabla W \) \( \sum_{i=1}^m \left[ \log p(z_t^t | x_i, \psi_t) - \log \nu V(x_i, z_t(z_i)) \right] \).
10. end for

Efficiency in Sample Complexity

Table: Convergence speed comparison in terms of number epoch on MNIST.

<table>
<thead>
<tr>
<th>Methods</th>
<th>log ( E(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVB (32-dim)</td>
<td>-84.0</td>
</tr>
<tr>
<td>AVB + AC (32-dim)</td>
<td>-9.9 [Mischler et al., 2017]</td>
</tr>
<tr>
<td>AVB + AC (32-dim)</td>
<td>-80.2 [Mischler et al., 2017]</td>
</tr>
<tr>
<td>DRAW + VGP</td>
<td>-79.9 [Tan et al., 2015]</td>
</tr>
<tr>
<td>VAE + IAF</td>
<td>-81.9 [Wongna et al., 2016]</td>
</tr>
<tr>
<td>VAE + NIF</td>
<td>-85.1 [Ouzda and Mohamed, 2018]</td>
</tr>
<tr>
<td>ConvVAE + HVI (T = 16)</td>
<td>-81.9 [Salimans et al., 2015]</td>
</tr>
<tr>
<td>VAE + HVI (T = 16)</td>
<td>-85.1 [Salimans et al., 2015]</td>
</tr>
</tbody>
</table>

Figure: Convergence speed comparison in terms of number epoch on MNIST.

Generative Ability

(a) Training data
(b) Random generation
(c) Reconstruction