Coupled Variational Bayes via Optimization Embedding

*Bo Dai[1,2], *Hanjun Dai[1], Niao He[3], Weiyang Liu[1], Zhen Liu[1], Jianshu Chen[4], Lin Xiao[5], Le Song[1,6]

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, $p_\theta(x,z) = p_\theta(x|z)p(z)$, with $x \in \mathbb{R}^d$ as the observations, $z \in \mathbb{R}^r$ latent variables, and θ as the parameters,

$$\log p_\theta(x) = \log \int p_\theta(x,z)dz \geq \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x,z) - \log q_\phi(z|x)\right].$$

Existing variational distributional families:
- **Reparametrized density:** Deep model parametrized simple distributions, *e.g.*,
$$q_\phi(z|x) := \mathcal{N}\left(z|\mu_{\phi_1}(x), \operatorname{diag}\left(\sigma^2_{\phi_2}(x)\right)\right)$$
with $\mu_{\phi_1}(x)$ and $\sigma_{\phi_2}(x)$ as neural networks.
Disadvantages: Simple densities restrict the approximation ability.
- **Tractable flow-based model:**
$$q_k(z|x) = q_0(z|x)\prod_{t=1}^T \left|\det \frac{\partial \mathcal{T}_t}{\partial z^t}\right|^{-1}$$
with $z^T = \mathcal{T}_T \circ \mathcal{T}_{T-1} \circ \ldots \circ \mathcal{T}_1(z^0)$ and $z_0 \sim q_0(z|x)$.
Disadvantages: **i),** $\{\mathcal{T}_t\}_{t=1}^T$ must invertible; **ii),** $\left|\det \frac{\partial \mathcal{T}_t}{\partial z^t}\right|^{-1}$ is expensive for general $\mathcal{T}_t$.

Key Component I: Primal-Dual view of ELBO

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

$$\max_{q(z|x)\in\mathcal{P}} \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x,z) - \log q_\phi(z|x)\right],$$

equivalently as

$$\min_{\nu\in\mathcal{H}_+} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{\xi\sim p_\xi(\cdot)}\left[\max_{z_{x,\xi}\in\mathbb{R}^r} \log p_\theta(x|z_{x,\xi}) - \log\nu(x,z_{x,\xi})\right] + \mathbb{E}_{z\sim p(z)}\left[\nu(x,z)\right]\right] - 1,$$

where $\mathcal{H}_+ = \left\{h : \mathbb{R}^d \times \mathbb{R}^r \to \mathbb{R}_+\right\}$, $p_\xi(\cdot)$ *denotes some simple distribution and the optimal* $\nu^*_\theta(x,z) = \frac{q^*_\theta(z|x)}{p(z)}$.

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

$$KL(q||p) = \left\langle q, \log\frac{q}{p}\right\rangle = \max_{\nu>0} \langle q, \log\nu\rangle - \langle p, \nu\rangle + 1,$$

and interchangeability principle [Dai et al. 2016]. □

- **Nonparametric representation of q:** We are able to represent the distributional operation on q by local variables $z_{x,\xi}$.
- **Efficient computation:** We can avoid the computation of log-det of Jacobian.

Key Component II: Optimization Embedding

By the Theorem, we can obtain the implicit nonparametric transformation from $(x,\xi) \in \mathbb{R}^d \times \Xi$ to $z_{x,\xi} \in \mathbb{R}^p$ as

$$z^*_{x,\xi;\theta} = \operatorname*{argmax}_{z_{x,\xi}\in\mathbb{R}^p} \log p_\theta(x|z_{x,\xi}) - \log\nu(x,z_{x,\xi}). \tag{1}$$

Optimization Embedding with Mirror Descent Algorithm (MDA):

$$z^t_{x,\xi;\theta} = \operatorname*{argmax}_{z\in\mathbb{R}^r} \left\langle z, \eta_t g\left(x, z^{t-1}_{x,\xi;\theta}\right)\right\rangle - D_\omega\left(z^{t-1}_{x,\xi;\theta}, z\right), = f^{-1}\left(\eta_t g\left(x, z^{t-1}_{x,\xi;\theta}\right) + f\left(z^{t-1}_{x,\xi;\theta}\right)\right).$$

where $g\left(x, z^{t-1}_{x,\xi;\theta}\right) = \nabla_z \log p_\theta\left(x|z^{t-1}_{x,\xi;\theta}\right) - \nabla_z \log\nu\left(x, z^{t-1}_{x,\xi;\theta}\right)$, $D_\omega(z_1,z_2) = \omega(z_2) - \left[\omega(z_1) + \langle\nabla\omega(z_1), z_2 - z_1\rangle\right]$ is the Bregman divergence, and $f(\cdot) = \nabla\omega(\cdot)$.

The MDA algorithm naturally establishes nonparametric function that maps from $\mathbb{R}^d \times \Xi$ to $\mathbb{R}^r$ to approximate the mapping defined by (1),

$$z^T_\theta(x,\xi) \approx z^*_{x,\xi;\theta} = \operatorname*{argmax}_{z_{x,\xi}\in\mathbb{R}^p} \log p_\theta(x|z_{x,\xi}) - \log\nu(x,z_{x,\xi}).$$

Unbiased Gradient Estimator: We have the surrogate as

$$\max_\theta \widetilde{L}(\theta) := \min_{\nu\in\mathcal{H}_+} \ell(\nu,\theta),$$

with $\ell(\nu,\theta) := \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{\xi\sim p(\xi)}\left[\log p_\theta\left(x|z^T_\theta(x,\xi)\right) - \log\nu\left(x, z^T_\theta(x,\xi)\right)\right] + \mathbb{E}_{z\sim p(z)}\left[\nu(x,z)\right]\right]$.
Denote $\nu^*_\theta(x,z) = \operatorname{argmin}_{\nu\in\mathcal{H}_+} \ell(\nu,\theta)$, we have the unbiased gradient estimator w.r.t. θ as

$$\frac{\partial\widetilde{L}(\theta)}{\partial\theta} = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{\xi\sim p(\xi)}\left[\frac{\partial\log p_\theta(x|z)}{\partial\theta}\Big|_{z=z^T_\theta(x,\xi)} + \frac{\partial\log p_\theta(x|z)}{\partial z}\Big|_{z=z^T_\theta(x,\xi)}\frac{\partial z^T_\theta(x,\xi)}{\partial\theta}\right] - \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{\xi\sim p(\xi)}\left[\frac{\partial\log\nu^*_\theta(x,z)}{\partial z}\Big|_{z=z^T_\theta(x,\xi)}\frac{\partial z^T_\theta(x,\xi)}{\partial\theta}\right]. \tag{2}$$

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

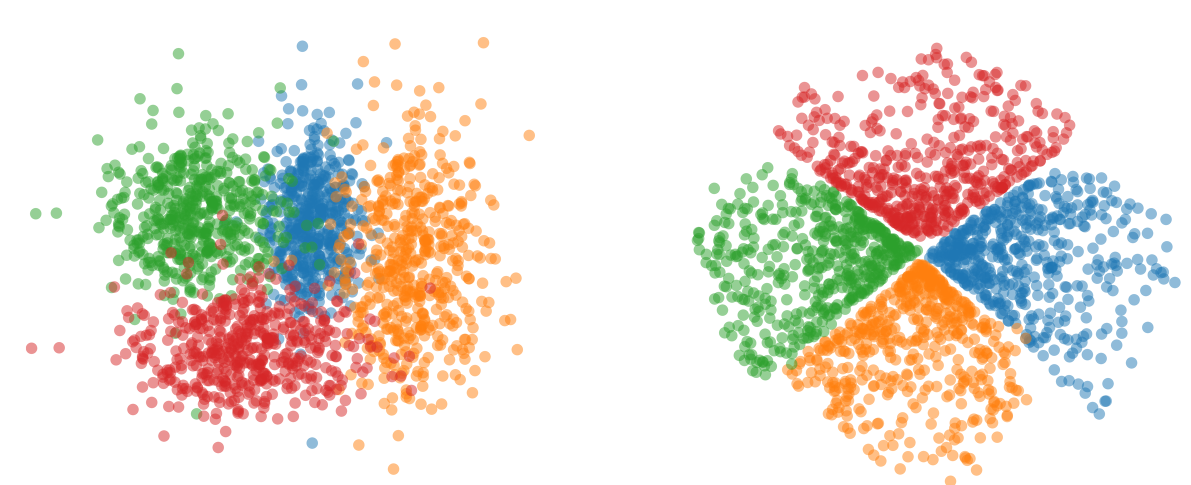
Optimization Embedding as Gradient Flow

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

$$\frac{\partial q_t(z|x)}{\partial t} = -\nabla\cdot\left(q_t(z|x)\,g_t(x,z)\right),$$

if $g_t(x,z) := \nabla_z \log p_\theta(x|z) - \nabla_z \log\nu^*_t(x,z)$ *with* $\nu^*_t(x,z) = \frac{q_t(z|x)}{p(z)}$. *Such process is a gradient flow of KL-divergence in the space of measures with 2-Wasserstein metric.*

Flexibility in Posterior Approximation



(1) vanilla VAE (2) CVB

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

Algorithm

Coupled Variational Bayes (CVB)

1: Initialize θ, V and W (the parameters of ν and z^0) randomly, set length of steps T and mirror function f.
Set $z^0(x,\xi) = h_W(x,\xi)$.
2: **for** iteration $k = 1, \ldots, K$ **do**
3: Sample mini-batch $\{x_i\}_{i=1}^m$ from dataset $\mathcal{D}$, $\{z_i\}_{i=1}^m$ from prior $p(z)$, and $\{\xi_i\}_{i=1}^m$ from $p(\xi)$.
4: **for** iteration $t = 1, \ldots, T$ **do**
5: Compute $z^t_\theta(x,\xi)$ for each pair of $\{x_i,\xi_i\}_{i=1}^m$.
6: **Descend** V with $\nabla_V \frac{1}{m}\sum_{i=1}^m\left[\nu_V(x_i,z_i) - \log\nu_V\left(x_i, z^t_\theta(x_i,\xi_i)\right)\right]$.
7: **end for**
8: **Ascend** θ by stochastic gradient (2).
9: **Ascend** W by $\nabla_W \frac{1}{m}\sum_{i=1}^m\left[\log p_\theta\left(x|z^T_\theta(x,\xi)\right) - \log\nu_V\left(x, z^T_\theta(x,\xi)\right)\right]$.
10: **end for**

Efficiency in Sample Complexity

Methods	$\log p(x) \approx$
CVB (8-dim)	-93.5
CVB (32-dim)	-84.0
AVB + AC (8-dim)	−89.6 [Mescheder et al., 2017]
AVB + AC (32-dim)	−80.2 [Mescheder et al., 2017]
DRAW + VGP	−79.9 [Tran et al., 2015]
VAE + IAF	−79.1 [Kingma et al., 2016]
VAE + NF ($T = 80$)	−85.1 [Rezende and Mohamed, 2015]
convVAE + HVI ($T = 16$)	−81.9 [Salimans et al., 2015]
VAE + HVI ($T = 16$)	−85.5 [Salimans et al., 2015]

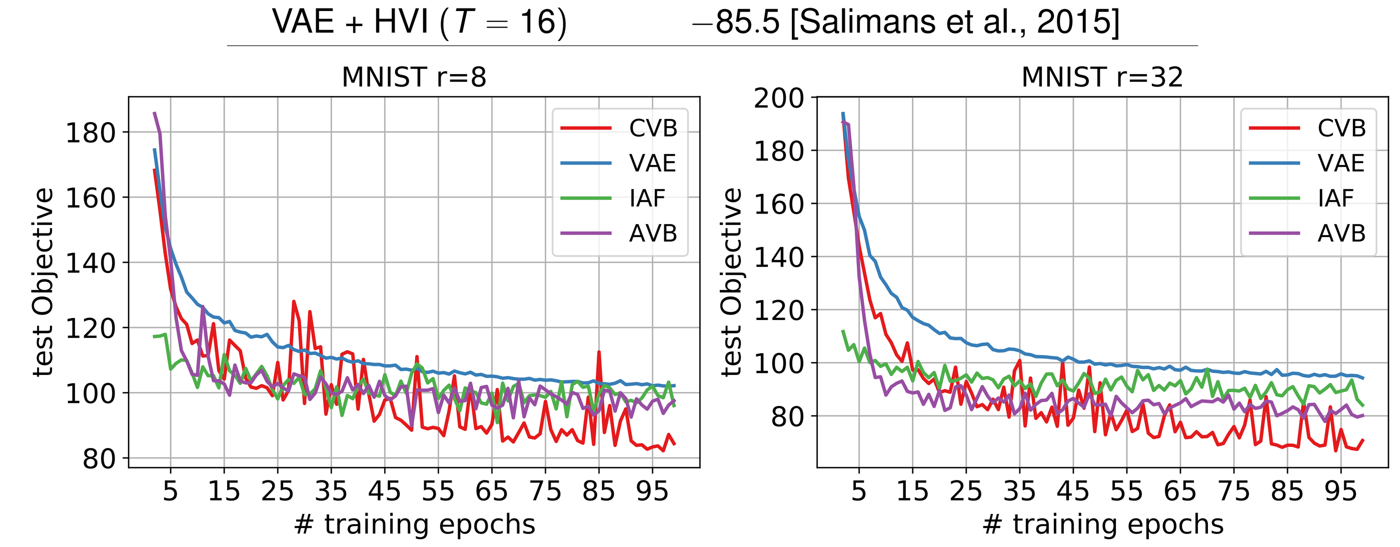
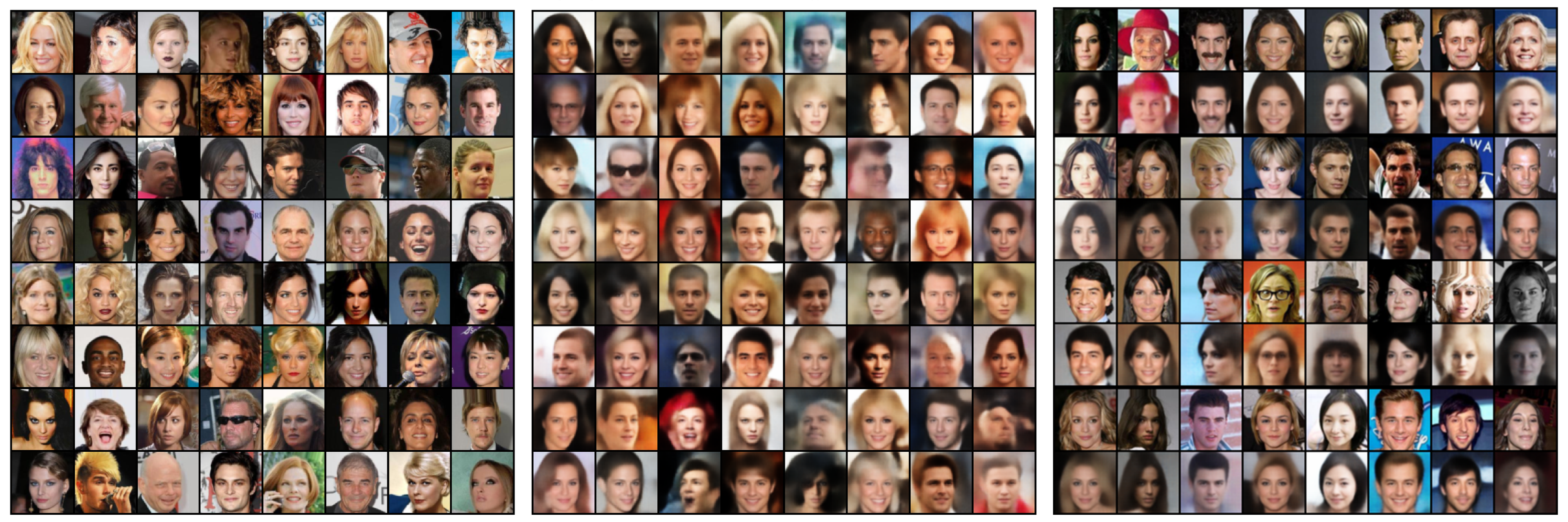


Figure: Convergence speed comparison in terms of number epoch on MNIST.

Generative Ability

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



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

[1]Georgia Institute of Technology, [2]Google Brain, [3]University of Illinois at Urbana Champaign, [4]Tencent AI Lab, [5]Microsoft Research, [6]Ant Financial

Mailto: bodai@google.com