Supplement Materials for Boosting Generative Models by Leveraging Cascaded Meta-Models

A Proof of the Decomposable Lower Bound

Theorem 1. *Let $m_i(h_{i-1}, h_i)(1 \leq i \leq n)$ be meta-models, $p_k(x, h_1, \cdots, h_k) = m_k(h_{k-1}, h_k) \prod_{i=1}^{k-1} m_i(h_{i-1}|h_i)$ be the model boosted from meta-models $m_i(1 \leq i \leq k)$, $q(h_1, \cdots, h_{k-1}|x) = \prod_{i=1}^{k-1} m_i(h_i|h_{i-1})$ be the approximate posterior, then we have:*

$$\mathbf{E}_D\left[\log p_k(x)\right] \geq \mathbf{E}_D \mathbf{E}_{q(h_1,\cdots,h_{k-1}|x)}\left[\log \frac{p_k(x, h_1, \cdots, h_{k-1})}{q(h_1, \cdots, h_{k-1}|x)}\right] = \sum_{i=1}^{k} \mathcal{L}_k, \tag{1}$$

where

$$\begin{aligned} \mathcal{L}_1 =& \mathbf{E}_D\left[\log m_1(x)\right] \\ \mathcal{L}_i =& \mathbf{E}_D \mathbf{E}_{q(h_{i-1}|x)}\left[\log m_i(h_{i-1})\right] \\ & - \mathbf{E}_D \mathbf{E}_{q(h_{i-1}|x)}\left[\log m_{i-1}(h_{i-1})\right] (2 \leq i \leq k). \end{aligned} \tag{2}$$

Proof. Using $q(h_1, \cdots, h_{k-1}|x)$ as the approximate posterior, we have a variational lower bound for $\log p_k(x)$:

$$\log p_k(x) \geq \mathbf{E}_{q(h_1,\cdots,h_{k-1}|x)}\left[\log \frac{p_k(x, h_1, \cdots, h_{k-1})}{q(h_1, \cdots, h_{k-1}|x)}\right],$$

with

$$\begin{aligned} \frac{p_k(x, h_1, \cdots, h_{k-1})}{q(h_1, \cdots, h_{k-1}|x)} =& \frac{m_k(h_{k-1}) \prod_{i=1}^{k-1} \frac{m_i(h_{i-1}, h_i)}{m_i(h_i)}}{\prod_{i=1}^{k-1} \frac{m_i(h_{i-1}, h_i)}{m_i(h_{i-1})}} \\ =& m_1(x) \prod_{i=1}^{k-1} \frac{m_{i+1}(h_i)}{m_i(h_i)}. \end{aligned}$$

Thus, the lower bound is equal to:

$$\begin{aligned} & \mathbf{E}_{q(h_1,\cdots,h_{k-1}|x)}\left\{\log\left[m_1(x) \prod_{i=1}^{k-1} \frac{m_{i+1}(h_i)}{m_i(h_i)}\right]\right\} \\ =& \log m_1(x) + \sum_{i=1}^{k-1} \mathbf{E}_{q(h_i|x)}\left[\log \frac{m_{i+1}(h_i)}{m_i(h_i)}\right] \\ =& \log m_1(x) + \sum_{i=1}^{k-1} \left\{\mathbf{E}_{q(h_i|x)}\left[\log m_{i+1}(h_i)\right]\right. \\ & \left. - \mathbf{E}_{q(h_i|x)}\left[\log m_i(h_i)\right]\right\}. \end{aligned}$$

Thus,

$$\begin{aligned} & \log p_k(x) \geq \log m_1(x) \\ & + \sum_{i=1}^{k-1} \left\{\mathbf{E}_{q(h_i|x)}\left[\log m_{i+1}(h_i)\right] - \mathbf{E}_{q(h_i|x)}\left[\log m_i(h_i)\right]\right\}. \end{aligned}$$

Take the expection with respect to dataset D, we have

$$\begin{aligned} & \mathbf{E}_D\left[\log p_k(x)\right] \geq \mathbf{E}_D\left[\log m_1(x)\right] + \sum_{i=1}^{k-1} \\ & \left\{\mathbf{E}_D \mathbf{E}_{q(h_i|x)}\left[\log m_{i+1}(h_i)\right] - \mathbf{E}_D \mathbf{E}_{q(h_i|x)}\left[\log m_i(h_i)\right]\right\} \\ & = \sum_{i=1}^{k} \mathcal{L}_k. \end{aligned}$$

□

B Proof of the Convergence

Theorem 2. *If m_k reaches the best global optimum, then $\sum_{i=k+1}^{j} \mathcal{L}_i \leq 0$, for any $j \geq k+1$.*

Theorem 3. *If m_k reaches the best global optimum, then $\mathbf{E}_D \mathbf{E}_{q(h_k|x)}\left[\log m_k(h_k)\right] = \mathbf{E}_{m_k(h_k)}\left[\log m_k(h_k)\right]$.*

Proof. Since m_k reaches the best global optimum, $m_k(h_{k-1})$ is exactly the marginal distribution of $p_D(x)q(h_{k-1}|x)$ and we have

$$\mathbf{E}_D \mathbf{E}_{q(h_{k-1}|x)}\left[\log m_k(h_{k-1})\right] = \mathbf{E}_{m_k(h_{k-1})}\left[\log m_k(h_{k-1})\right].$$

For $i \geq k+1$, we have

$$\begin{aligned} \mathcal{L}_i =& \mathbf{E}_{m_k(h_{k-1})} \mathbf{E}_{q(h_{i-1}|h_{k-1})}\left[\log m_i(h_{i-1})\right] \\ & - \mathbf{E}_{m_k(h_{k-1})} \mathbf{E}_{q(h_{i-1}|h_{k-1})}\left[\log m_{i-1}(h_{i-1})\right]. \end{aligned}$$

Given $j \geq k+1$, let

$$p_j(h_{k-1}, \cdots, h_j) = m_j(h_{j-1}, h_j) \prod_{i=k}^{j-1} m_i(h_{i-1}|h_i).$$

and let

$$q(h_k, \cdots, h_{j-1}|h_{k-1}) = \prod_{i=k}^{j-1} m_i(h_i|h_{i-1})$$

be the approximate posterior of $p_j(h_{k-1}, \cdots, h_{j-1})$.

According to Theorem 1, we have

$$\begin{aligned}&\mathbf{E}_{m_k(h_{k-1})}\left[\log p_j(h_{k-1})\right]\\ \geq&\mathbf{E}_{m_k(h_{k-1})}\left[\log m_k(h_{k-1})\right]+\sum_{i=k+1}^{j}\mathcal{L}_i.\end{aligned}$$

Since

$$\mathbf{E}_{m_k(h_{k-1})}\left[\log p_j(h_{k-1})\right]\leq\mathbf{E}_{m_k(h_{k-1})}\left[\log m_k(h_{k-1})\right],$$

we have $\sum_{i=k+1}^{j}\mathcal{L}_i\leq 0$. Besides, we also have

$$\begin{aligned}&\mathbf{E}_D\mathbf{E}_{q(h_k|x)}\left[\log m_k(h_k)\right]\\ &\quad=\mathbf{E}_{m_k(h_{k-1})}\mathbf{E}_{q(h_k|h_{k-1})}\left[\log m_k(h_k)\right]\\ &\quad=\mathbf{E}_{m_k(h_k)}\left[\log m_k(h_k)\right].\end{aligned}$$

□

C Semi-supervised Boosting

Our meta-algorithmic framework can be extended to semi-supervised boosting. Suppose we have a large amount of labelled data and a small amount of unlabelled data, where the unlabelled dataset is denoted by D_u and the labelled dataset is denoted by D_l. The task is to learn a conditional distribution from these data.

To solve this task, we introduce a latent class variable y. The first $n-1$ meta-models doesn't have the latent class variable y; they are normal hidden variable models. The latent class variable y only appears in the last meta-model. Thus, the joint distribution of the boosted model become

$$p_n(x,y,h_1,\cdots,h_n)=m_n(h_{n-1},y,h_n)\prod_{i=1}^{n-1}m_i(h_{i-1}|h_i). \tag{3}$$

Similarly, the approximation of the posterior distribution become

$$q(y,h_1,\cdots,h_n|x)=m_n(y,h_n|h_{n-1})\prod_{i=1}^{n-1}m_i(h_i|h_{i-1}). \tag{4}$$

For unlabelled data, we have the following decomposable lower bound,

$$\mathbf{E}_{D_u}\left[\log p_n(x)\right]\geq\sum_{i=1}^{n}\mathcal{L}_i^u, \tag{5}$$

where each decomposed term $\mathcal{L}_i^u$ is

$$\begin{aligned}\mathcal{L}_1^u=&\mathbf{E}_{D_u}\left[\log m_1(x)\right]\\ \mathcal{L}_i^u=&\mathbf{E}_{D_u}\mathbf{E}_{q(h_{i-1}|x)}\left[\log m_i(h_{i-1})\right]\\ &-\mathbf{E}_{D_u}\mathbf{E}_{q(h_{i-1}|x)}\left[\log m_{i-1}(h_{i-1})\right](2\leq i\leq n).\end{aligned} \tag{6}$$

We see that these terms have no difference from the decomposed terms listed in Theorem 1, except that the expectations of marginal likelihoods are taken on unlabelled dataset D_u. The proof for Equation 5 is given in C.2.

For labelled data, we have the following decomposable lower bound,

$$\mathbf{E}_{D_l}\left[\log p_n(x,y)\right]\geq\sum_{i=1}^{n}\mathcal{L}_i^l, \tag{7}$$

where the first $n-1$ decomposed terms are

$$\begin{aligned}\mathcal{L}_1^l=&\mathbf{E}_{D_l}\left[\log m_1(x)\right]\\ \mathcal{L}_i^l=&\mathbf{E}_{D_l}\mathbf{E}_{q(h_{i-1}|x)}\left[\log m_i(h_{i-1})\right]\\ &-\mathbf{E}_{D_l}\mathbf{E}_{q(h_{i-1}|x)}\left[\log m_{i-1}(h_{i-1})\right](2\leq i\leq n-1),\end{aligned} \tag{8}$$

which have no difference from the decomposed terms listed in Theorem 1, except that the expectations of marginal likelihood are taken on labelled dataset D_l, and the last term is

$$\begin{aligned}\mathcal{L}_n^l=&\mathbf{E}_{D_l}\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_n(h_{n-1},y)\right]\\ &-\mathbf{E}_{D_l}\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_{n-1}(h_{n-1})\right],\end{aligned} \tag{9}$$

which has a slight difference from the last term listed in Theorem 1: the first part of $\mathcal{L}_n^l$ is the marginal likelihood of the ensemble of the observable variable h_{n-1} and the latent class variable y. It is in $\mathcal{L}_n^l$ where the latent class variable y is considered. The proof for Equation 7 is given in C.3.

We use a weighted sum of Equation 5 and Equation 7 to get a lower bound for the entire dataset:

$$\begin{aligned}&\alpha\mathbf{E}_{D_u}\left[\log p_n(x)\right]+\beta\mathbf{E}_{D_l}\left[\log p_n(x,y)\right]\\ \geq&\alpha\sum_{i=1}^{n}\mathcal{L}_i^u+\beta\sum_{i=1}^{n}\mathcal{L}_i^l\\ =&\sum_{i=1}^{n}\left[\alpha\mathcal{L}_i^u+\beta\mathcal{L}_i^l\right]\\ =&\sum_{i=1}^{n}\mathcal{J}_i,\end{aligned} \tag{10}$$

where α and β are the weights and $\alpha+\beta=1$. The lower bound is also decomposed to n terms. If we let D denote the weighted combination of D_u and unlabelled version of D_l (i.e., the empirical distribution of D is the weighted summation of p_{D_u} and p_{D_l}: $p_D(x)=\alpha p_{D_u}(x)+\beta p_{D_l}(x)$), we can get

$$\begin{aligned}\mathcal{J}_1=&\mathbf{E}_D\left[\log m_1(x)\right]\\ \mathcal{J}_i=&\mathbf{E}_D\mathbf{E}_{q(h_{i-1}|x)}\left[\log m_i(h_{i-1})\right]\\ &-\mathbf{E}_D\mathbf{E}_{q(h_{i-1}|x)}\left[\log m_{i-1}(h_{i-1})\right](2\leq i\leq n-1),\end{aligned} \tag{11}$$

which also have no difference from these of the lower bound in Theorem 1, and

$$\begin{aligned}\mathcal{J}_n=&\alpha\mathbf{E}_{D_u}\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_n(h_{n-1})\right]\\ &+\beta\mathbf{E}_{D_l}\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_n(h_{n-1},y)\right]\\ &-\mathbf{E}_D\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_{n-1}(h_{n-1})\right].\end{aligned} \tag{12}$$

Since the first $n-1$ terms of the lower bound also have no difference from these of the lower bound in Theorem 1, we can train the first $n-1$ meta-models by directly using our algorithm of boosting generative models. For the last meta-model,

Algorithm 1 Semi-Supervised Boosting

1: **Input:** unlabelled dataset D_u; labelled dataset D_l; number of meta-models n
2: Train first $n-1$ meta-models using our algorithm of boosting generative models
3: Train the last meta-model m_n by optimizing $\mathcal{J}_n$
4: Build the boosted model $p_n(x, y, h_1, \cdots, h_n) = m_n(h_{n-1}, y, h_n) \prod_{i=1}^{n-1} m_i(h_{i-1}|h_i)$
5: Build the approximate posterior $q(y, h_1, \cdots, h_n|x) = m_n(y, h_n|h_{n-1}) \prod_{i=1}^{n-1} m_i(h_i|h_{i-1})$
6: Return $p_n(x, y, h_1, \cdots, h_n)$, $q(y, h_1, \cdots, h_n|x)$

we train it by optimizing $\mathcal{J}_n$: only tune the parameters of m_n to make $\mathcal{J}_n$ grow. The training algorithm is summarized in Algorithm 1.

It is worth note that, when we restrict the number of meta-models to 2, the boosted model is exactly the stacked generative semi-supervised model [Kingma *et al.*, 2014].

We don't introduce a classification loss $\mathbf{E}_{D_l}\left[-\mathbf{log}q(y|x)\right]$ [Kingma *et al.*, 2014] for the boosted model. In our experiments, we will show that the performances of models with the loss term and models without the loss term are comparable on the classification task. In fact, by well balancing α and β, the boosted-model can have the ability for classification, which reveals that the classification task can be implemented entirely by generative methods.

C.1 Experiments on Semi-supervised Boosting

This experiment is designed to show that deep generative models can perform well on the classification task without the help of any classification loss. We compare our methods (without classification loss) with the methods in Kingma *et al.* [2014] (with classification loss). The models covered in our experiments are listed in Table 1. M_1 and M_2 are models in our methods with no classification loss; M_3 and M_4 are models in the methods of Kingma *et al.* [2014] with classification loss.

	has classification loss?	#meta-models
M_1	Y	1
M_2	Y	2
M_3	N	1
M_4	N	2

Table 1: The boosted models covered in our experiments of semi-supervised learning.

We do experiments under different amount of labeled data: we randomly pick 100, 600, and 1000 and 3000 labelled data from mnist training data, where each class has the same number of labelled data. The results are shown in Table 2. The classification accuracy for models with and without the classification loss are comparable. Under some cases, the models without loss terms can outperform models with loss terms (M_1 outperforms M_3 on any number of labelled data, while they have the same number of meta-models). This result suggests us that the generative model can learn how to classify without telling it how to classify.

$\lvert D_l\rvert$	M_1	M_2	M_3	M_4
100	94.32	95.20	88.03	96.67
600	96.04	95.74	95.06	97.41
1000	97.38	96.10	96.40	97.60
3000	97.52	96.57	96.08	97.82

Table 2: The classification accuracy on mnist test dataset for models with and without the classification loss. $|D_l|$ is the number of labelled data.

C.2 Proof of the Decomposable Lower Bound for Unlabelled Data

Proof. If we view h_n and y as a whole H_n, the joint distribution of the boosted model can be written as

$$p_n(x, h_1, \cdots, H_n) = m_n(h_{n-1}, H_n) \prod_{i=1}^{n-1} m_i(h_{i-1}|h_i),$$

and the approximation of the posterior distribution can be written as

$$q(h_1, \cdots, H_n|x) = m_n(H_n|h_{n-1}) \prod_{i=1}^{n-1} m_i(h_i|h_{i-1}).$$

Now, we can directly leverage Theorem 1 to get the decomposable lower bound:

$$\mathbf{E}_{D_u}\left[\log p_n(x)\right] \geq \sum_{i=1}^{n} \mathcal{L}_i^u, \tag{13}$$

where each decomposed term $\mathcal{L}_i^u$ is

$$\begin{aligned}\mathcal{L}_1^u =& \mathbf{E}_{D_u}\left[\log m_1(x)\right] \\ \mathcal{L}_i^u =& \mathbf{E}_{D_u}\mathbf{E}_{q(h_{i-1}|x)}\left[\log m_i(h_{i-1})\right] \\ & - \mathbf{E}_{D_u}\mathbf{E}_{q(h_{i-1}|x)}\left[\log m_{i-1}(h_{i-1})\right] (2 \leq i \leq n).\end{aligned} \tag{14}$$

□

C.3 Proof of the Decomposable Lower Bound for Labelled Data

Proof. The marginal likelihood of labelled data can be broken down into two parts:

$$\mathbf{E}_{D_l}\left[\log p_n(x, y)\right] = \mathbf{E}_{D_l}\left[\log p_n(x|y)\right] + \mathbf{E}_{D_l}\left[\log m_n(y)\right].$$

We derive a lower bound for the first part $\mathbf{E}_{D_l}\left[\log p_n(x|y)\right]$. Since the boosted model conditioned on y is

$$p_n(x, h_1, \cdots, h_n|y) = m_n(h_{n-1}, h_n|y) \prod_{i=1}^{n-1} m_i(h_{i-1}|h_i),$$

and the approximation of the posterior conditioned on y is

$$q(h_1, \cdots, h_n|x, y) = m_n(h_n|h_{n-1}, y) \prod_{i=1}^{n-1} m_i(h_i|h_{i-1}).$$

By leveraging Theorem 1, we can derive a lower bound for $\mathbf{E}_{D_l}\left[\log p_n(x|y)\right]$:

$$\mathbf{E}_{D_l}\left[\log p_n(x|y)\right] \geq \sum_{i=1}^{n-1} \mathcal{L}_i^l + \hat{\mathcal{L}}_n^l,$$

where

$$\begin{aligned}\mathcal{L}_1^l =& \mathbf{E}_{D_l}\left[\log m_1(x)\right] \\ \mathcal{L}_i^l =& \mathbf{E}_{D_l}\mathbf{E}_{q(h_{i-1}|x)}\left[\log m_i(h_{i-1})\right] \\ & - \mathbf{E}_{D_l}\mathbf{E}_{q(h_{i-1}|x)}\left[\log m_{i-1}(h_{i-1})\right] (2 \leq i \leq n-1),\end{aligned}$$

and

$$\begin{aligned}\hat{\mathcal{L}}_n^l =& \mathbf{E}_{D_l}\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_n(h_{n-1}|y)\right] \\ & - \mathbf{E}_{D_l}\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_{n-1}(h_{n-1})\right].\end{aligned}$$

Then we have

$$\begin{aligned}\mathbf{E}_{D_l}\left[\log p_n(x,y)\right] =& \mathbf{E}_{D_l}\left[\log p_n(x|y)\right] + \mathbf{E}_{D_l}\left[\log m_n(y)\right] \\ \geq & \sum_{i=1}^{n-1} \mathcal{L}_i^l + \hat{\mathcal{L}}_n^l + \mathbf{E}_{D_l}\left[\log m_n(y)\right],\end{aligned}$$

and

$$\begin{aligned}& \hat{\mathcal{L}}_n^l + \mathbf{E}_{D_l}\left[\log m_n(y)\right] \\ =& \mathbf{E}_{D_l}\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_n(h_{n-1}|y)\right] + \mathbf{E}_{D_l}\left[\log m_n(y)\right] \\ & - \mathbf{E}_{D_l}\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_{n-1}(h_{n-1})\right] \\ =& \mathbf{E}_{D_l}\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_n(h_{n-1}, y)\right] \\ & - \mathbf{E}_{D_l}\mathbf{E}_{q(h_{n-1}|x)}\left[\log m_{n-1}(h_{n-1})\right] \\ =& \mathcal{L}_n^l.\end{aligned}$$

Thus, we have $\mathbf{E}_{D_l}\left[\log p_n(x,y)\right] \geq \sum_{i=1}^{n} \mathcal{L}_i^l$.

□

D Architectures of Meta-Models

The architectures of VAEs, ConvVAEs, IWAEs and LVAEs are given in this part.

D.1 Architectures of VAEs

All VAEs have two deterministic hidden layers for both generation, and inference and we add batch normalization layers [Ioffe and Szegedy, 2015; Sønderby *et al.*, 2016] after deterministic hidden layers. The dimension of deterministic hidden layers is set to 500 and 2500, and the dimension of stochastic hidden variables is set to 20 and 100, for experiments on mnist and celebA respectively.

D.2 Architectures of ConvVAEs

The ConvVAE has one 500-dimensional deterministic hidden layer and one 50-dimensional stochastic hidden variable, with four additional convolutional layers [LeCun *et al.*, 1998]. All convolutional layers have a kernel size of 4×4 and a stride of 2. Their channels are 32, 64, 128 and 256 respectively. We add batch normalization layers after deterministic hidden layers.

D.3 Architectures of IWAEs

The IWAE has two 500-dimensional deterministic hidden layers and one 50-dimensional stochastic hidden variable. The number of importance sampling is set to 5 and 10.

D.4 Architectures of LVAEs

The LVAE has four 1000-dimensional deterministic hidden layers and two 30-dimensional stochastic hidden variables. We add batch normalization layers after deterministic hidden layers.

References

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[Kingma *et al.*, 2014] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Sønderby *et al.*, 2016] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.