A Proofs

Proposition 1. *If $X_i - Y_i$ has the expected cumulant-generation function $\bar{\psi}_i(\lambda)$ and the inverse of the convex conjugate of $\bar{\psi}$ denoted by $\bar{\psi}^{*-1}$ exists, then the improved bound utilizing I_{NCE} can be expressed as follows,*

$$\mathbb{E}[X_i - Y_i] \leq \bar{\psi}^{*-1}(I_{NCE})$$

Proof. The proof is straightforward and based on original generalization bound in **??**. The result is obtained by following the steps of Theorem 2 [5] and replacing $I(X_i; Y_i)$ with I_{NCE} in the last step as follows,

$$\begin{aligned}
\mathbb{E}[X_i - Y_i] &\leq \inf_{\lambda \in [0, \min_i b_i)} \frac{\bar{\psi}(\lambda) + I(X_i; Y_i)}{\lambda} \\
\mathbb{E}[X_i - Y_i] &\leq \inf_{\lambda \in [0, \min_i b_i)} \frac{\bar{\psi}(\lambda) + I_{NCE}}{\lambda}; \quad since I_{NCE} \leq I(X_i; Y_i) \\
&= \bar{\psi}^{*-1}(I_{NCE})
\end{aligned}$$

This completes the proof. □

Proposition 2. *Given a function $f(w) : w \to \mathbb{R}$, the operator $\mathcal{T} f(w) = \log \sum_w \exp f(w)$ forms a contraction on $f(w)$.*

Proof. Let us first define a norm on the functions $||f_1(w) - f_2(w)|| \equiv \max_w |f_1(w) - f_2(w)|$. Suppose $\epsilon = ||f_1(w) - f_2(w)||$,

$$\begin{aligned}
\log \sum_w \exp\left(f_1(w)\right) &\leq \log \sum_w \exp\left(f_2(w) + \epsilon\right) \\
= \log \sum_w \exp\left(f_1(w)\right) &\leq \log \exp\left(\epsilon\right) \sum_w \exp\left(f_2(w)\right) \\
= \log \sum_{w=1} \exp\left(f_1(w)\right) &\leq \epsilon + \log \sum_w \exp\left(f_2(w)\right) \\
= \log \sum_w \exp\left(f_1(w)\right) - \log \sum_w \exp\left(f_2(w)\right) &\leq ||f_1(w) - f_2(w)|| \qquad (16)
\end{aligned}$$

Similarly, using ϵ with $\log \sum_w \exp\left(f_1(w)\right)$,

$$\begin{aligned}
\log \sum_w \exp\left(f_1(w) + \epsilon\right) &\geq \log \sum_w \exp\left(f_2(w)\right) \\
= \log \exp\left(\epsilon\right) \sum_w \exp\left(f_1(w)\right) &\geq \log \sum_w \exp\left(f_2(w)\right) \\
= \epsilon + \log \sum_w \exp\left(f_1(w)\right) &\geq \log \sum_w \exp\left(f_2(w)\right) \\
= ||f_1(w) - f_2(w)|| \geq \log \sum_w \exp\left(f_2(w)\right) - \log \sum_w \exp\left(f_1(w)\right) &\qquad (17)
\end{aligned}$$

Results in Equation 16 and Equation 17 prove that $\mathcal{T}$ is a contraction. □

Proposition 3. *If the operator $\mathcal{T} f(w) = \log \sum_w \exp f(w)$ is a contraction mapping $\forall w$, then the nonlinear interpolation bound I_{IN} can be further simplified as expressed in Equation 13.*

$$\begin{aligned}
I_{IN} \geq I_N \triangleq 1 + \log \sum_{i=1}^{K} \exp \mathbb{E}_{p(x_{1:K})p(y|x_i)} [\log \frac{e^{f(x_i, y)}}{\gamma m(y; x_{1:K}) + (1-\gamma) q(y)}] \\
\log \sum_{i=1}^{K} \exp \mathbb{E}_{p(x_{1:K})p(y)} [\log \frac{e^{f(x_i, y)}}{\gamma m(y; x_{1:K}) + (1-\gamma) q(y)}]
\end{aligned}$$

Proof. From Equation 7, we have the following,

$$\begin{aligned}
I_{IN} \triangleq 1 + \mathbb{E}_{p(x_{1:K})p(y|x_1)} [\log \frac{e^{f(x_1, y)}}{\gamma m(y; x_{1:K}) + (1-\gamma) q(y)}] \\
- \mathbb{E}_{p(x_{1:K})p(y)} [\log \frac{e^{f(x_1, y)}}{\gamma m(y; x_{1:K}) + (1-\gamma) q(y)}]
\end{aligned}$$

For notational convenience, we relabel the two interpolation terms as follows,

$$\mathcal{P}(x) = \mathbb{E}_{p(x_{1:K})p(y|x_1)}[\log \frac{e^{f(x_1,y)}}{\gamma m(y; x_{1:K}) + (1-\gamma)q(y)}]$$
$$\mathcal{Q}(x) = \mathbb{E}_{p(x_{1:K})p(y)}[\log \frac{e^{f(x_1,y)}}{\gamma m(y; x_{1:K}) + (1-\gamma)q(y)}]$$

Using the result from Proposition 2 and applying $\mathcal{T}$ to the two interpolation terms in the expressions gives us the following,

$$\mathcal{TP}(x) - \mathcal{TQ}(x) \leq \max_x |\mathcal{P}(x) - \mathcal{Q}(x)|$$
$$\mathcal{TP}(x) - \mathcal{TQ}(x) \leq \mathcal{P}(x) - \mathcal{Q}(x)$$

Using the above result in I_{IN} yields the lower bound I_N and completes the proof. □

Proposition 4. *Since $I_{NCE} \geq I_N$ with high probability, the generalization bound in Proposition 1 can be simplified as follows,*

$$\mathbb{E}[X_i - Y_i] \leq \bar{\psi}^{*-1}(I_N) \leq \bar{\psi}^{*-1}(I_{NCE})$$

Proof. We follow the steps of Proposition 1 to obtain the lower bound with I_{NCE} and replace it with I_N as follows,

$$\mathbb{E}[X_i - Y_i] \leq \inf_{\lambda \in [0, \min_i b_i)} \frac{\bar{\psi}(\lambda) + I(X_i; Y_i)}{\lambda}$$
$$\mathbb{E}[X_i - Y_i] \leq \inf_{\lambda \in [0, \min_i b_i)} \frac{\bar{\psi}(\lambda) + I_{NCE}}{\lambda}; \quad since I_{NCE} \leq I(X_i; Y_i)$$
$$\mathbb{E}[X_i - Y_i] \leq \inf_{\lambda \in [0, \min_i b_i)} \frac{\bar{\psi}(\lambda) + I_N}{\lambda}; \quad since I_N \leq I_{NCE}$$
$$= \bar{\psi}^{*-1}(I_N)$$

This completes the proof.

□

B Additional Results

B.1 ResNet-18

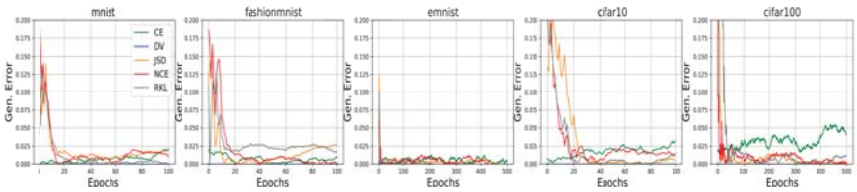

Figure 3: Gen. Error for variational objectives on standard and large-scale benchmarks for the ResNet-18 model. ϕ-divergence measures demonstrate improved stability and generalization in comparison to conventional InfoNCE and fully-supervised CE which depict high bias. Conventional DV, on the other hand, presents high variance and unstable convergence as a result of exponential estimates of the partition function.

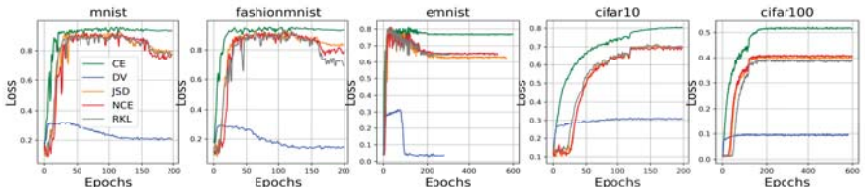

Figure 4: Comparison of Top1 classification accuracies for variational objectives on standard and large-scale benchmarks for the ResNet-18 model. While fully-supervised CE demonstrates best performance, InfoNCE and ϕ-divergence measures depict comparably accurate performance. Conventional DV, on the other hand, demonstrates a failure to learn.

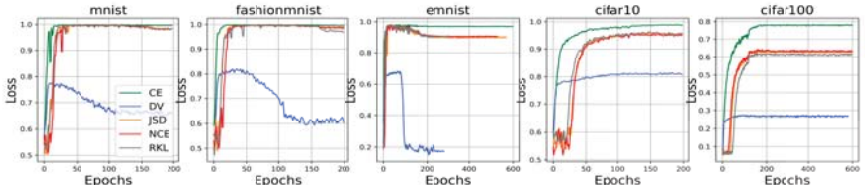

Figure 5: Comparison of Top5 classification accuracies for variational objectives on standard and large-scale benchmarks for the ResNet-18 model. Performance is analogous to Top1 counterpart with variational bounds demonstrating improved performance on standard mnist and fashionmnist benchmarks.

B.2 ResNet-34

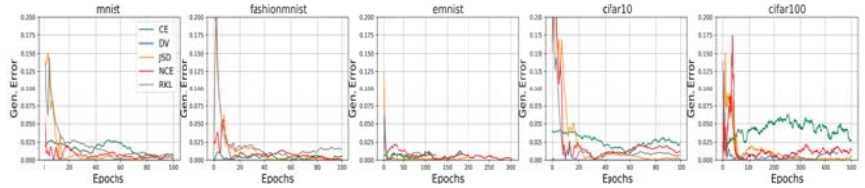

Figure 6: Gen. Error for variational objectives on standard and large-scale benchmarks for the ResNet-34 model. ϕ-divergence measures scale to well to the deeper ResNet-34 model and maintain minimum bias in generalization.

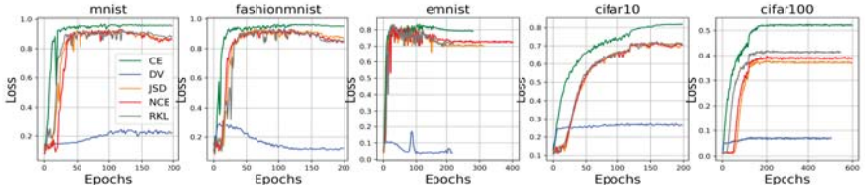

Figure 7: Comparison of Top1 classification accuracies for variational objectives on standard and large-scale benchmarks for the ResNet-34 model. Bounds tend to overfit in the case of EMNIST dataset as a result of large (26) number of classes.

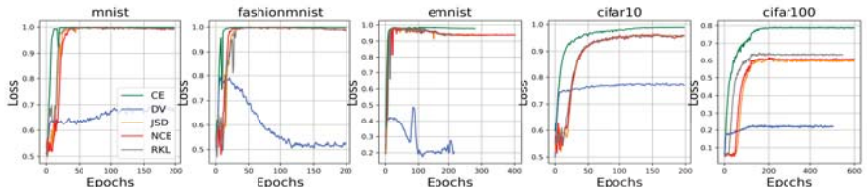

Figure 8: Comparison of Top5 classification accuracies for variational objectives on standard and large-scale benchmarks for the ResNet-34 model. Variational bounds demonstrate comparable performance to fully-supervised CE on MNIST, FashionMNIST and EMNIST benchmarks.

C Implementation Details

C.1 Note on Experiments

The experimental setup is based on the instance discrimination framework of [10] which aims to extract rich feature representations from input images without parameteric estimations. We extend this setup to variational bounds for the purpose of assessing generalization in learning. While training on standard benchmarks consisted of 200 epochs, large-scale benchmarks make use of 600 epochs. Hyperparameters corresponding to each model were fixed and were only changed for the purpose of tuning DV objective. DV does not demonstrate learning on datasets as a result of exponential results which inject instability in gradients. While measures such as gradient clipping were used, they did little to improve the stability of DV. A more suitable approach o stabilize the gradient in DV could be by normalizing the objective with the running mean of the gradient as presented in [13]. The normalization constant acts as the pseudo partition function which increases the stability of updates. However, in order to maintain uniformity in experiments and solely evaluate the role of variation bound, we do not make use of this trick.

C.2 Hyperparameters

Hyperparameters for ResNet-18 and ResNet-34 are kept same for all tasks. While CE, JSD, InfoNCE, and RKL make use of the same set of hyperparameters, DV makes use of a separate learning rate. The following table presents hyperparameters used for experiments-

Hyperparameter	Value
optimizer	SGD
momentum	0.9
weight decay	5e-4
learning rate	0.03 (0.007 for DV)
learning schedule	decay by 10 every 30 epochs
batch size	128
number of epochs	200 (600 for large datasets)
number of runs	1
NCE number of negative samples	512
NCE temperature	0.07
embedding dimension	128
NCE momentum	0.5