Pseudo-Encoded Stochastic Variational Inference

Amir Zadeh [1]   **Simon Hessner** [1]   **Yao-Chong Lim** [2]   **Louis-Philippe Morency** [1]

Abstract

Posterior inference in directed graphical models is commonly done using a probabilistic encoder (a.k.a inference model) conditioned on the input. Often this inference model is trained jointly with the probabilistic decoder (a.k.a generator model). If probabilistic encoder encounters complexities during training (e.g. suboptimal complxity or parameterization), then learning reaches a suboptimal objective; a phenomena commonly called inference suboptimality (Cremer et al., 2018). In Variational Inference (VI)(Jordan et al., 1999), optimizing the ELBo using Stochastic Variational Inference (SVI) (Rezende et al., 2014) can eliminate the inference suboptimality (as demonstrated in this paper), however, this solution comes at a substantial computational cost when inference needs to be done on new data points. Essentially, a long sequential chain of gradient updates is required to fully optimize approximate posteriors. In this paper, we present an approach called Pseudo-Encoded Stochastic Variational Inference (PE-SVI), to reduce the inference complexity of SVI during test time. Our approach relies on finding a suitable initial start point for gradient operations, which naturally reduces the required gradient steps. Furthermore, this initialization allows for adopting larger step sizes (compared to random initialization used in SVI), which further reduces the inference time complexity. PE-SVI reaches the same ELBo objective as SVI using less than one percent of required steps, on average.

1. Introduction

Training directed graphical models using Variational Inference (VI) has a long history in machine learning research (Jordan et al., 1999) . Commonly, inference is done

---

[1]LTI, SCS, Carnegie Mellon University [2]SCS, Carnegie Mellon University. Correspondence to: Amir Zadeh <abagherz@cs.cmu.edu>.

using probabilistic inference models (Dayan et al., 1995) such as a probabilistic encoder in VAE (Kingma & Welling, 2013). Using a parameteric model to perform inference allows for fast inference given new datapoints. However, if inference network encounters difficulties, then maximization of ELBo is done suboptimally (Cremer et al., 2018). Previous works have attempted to mitigate the inference suboptimality using fine-tuning (Hjelm et al., 2016), ladder-based models (Sønderby et al., 2016) and Hessian-based models (Kim et al., 2018). While these attempts have been very successful in dealing with numerical instabilities, inference suboptimality due to limited inference model capacity is intertwined with the nature of inference models CITE. Alternatively, to avoid this inference suboptimality altogether, as shown in this paper, one can rely on Stochastic Variational Inference (SVI) using free-form posterior parameterization and mean-field approximation (Rezende et al., 2014). However, using SVI, inference for new datapoint requires a long gradient (or somewhat faster alternative meta-gradient approaches) update chain, which makes the inference suffer heavily during test time. Essentially, parameters of approxiamte posteriors are initialized randomly and updated iteratively until ELBo maximization objective is reached.

In this paper, we assume the following separation about the inference process of SVI: 1) a suboptimal initial inference that a reasonably parameterized inference model can reach, 2) subsequent gradient-based updates to reach full ELBo maximization. Using the above assumption, we reach at a simple-yet-elegant framework called Pseudo-Encoded Stochastic Variational Inference (PE-SVI): a framework for test-time inference speed-up of SVI. The learning process is separated in three parts: (a) *Early Decoder Training*: which trains a decoder using SVI to maximize the lower-bound of likelihood using tractable easy-to-sample approximate posteriors. (b) *Deferred Encoder Training*: After the decoder and approximate posterior parameters are fully learned over the train set, a *pseudo-encoder* is trained in a supervised fashion between input data points and their respective approximate posterior parameters. *Pace Adjustment*: After initial approximate posterior parameter estimation using the trained encoder, the step size can be increased and tuned for fast convergence. Such large step sizes are often detrimental to SVI if approximate posterior parameters are initialized randomly.

The following summarizes, contributions and findings of this paper:

- We present a speed-up framework for test-time Stochastic Variational Inference (SVI), called Pseudo-Encoded Stochastic Variational Inference (PE-SVI). PE-SVI is easy to implement and does not require complex or costly calculations during train time (e.g. Hessian calculations (Kim et al., 2018)).

- PE-SVI is able to reach the same ELBo as SVI, with a fraction of the required steps. In our experiments over publicly available datasets, PE-SVI reaches similar performance as SVI in an average of 15.2 gradient updates, while SVI takes substantially larger number of steps with an average of 1826.1.

- To our surprise, ELBo loss achieved using PE-SVI's pseudo-encoder without any gradient steps in majority of times is better than end-to-end training of both encoder and decoder for VI (i.e. VAE). In simple terms, our experiments controversially hint that it is better to train the decoder first and subsequently the encoder, as opposed to training both end-to-end. This is further discussed in Section 5.

2. Background and Related Works

In this section we first start with the background required for VI and SVI. We subsequently discuss the comparison between our approach and previous methods for improving SVI inference complexity.

2.1. Variational Inference

Let samples drawn as $(z, x) \sim p(z)p(x|z)$ form a dataset $S = \{(z_i, x_i)\}_{i=1}^{|S|}$. x_i, z_i are regarded as observed and latent variables. Unfortunately z_i, being the latent space generating the data x_i, is not observable. Therefore, MLE on the joint distribution is not possible. Considering a parametric distribution with parameters θ, the marginal likelihood can be written as:

$$
\begin{aligned}
\mathcal{L}^{(i)}(\theta) = \log \int p_\theta(z, x_i) \, dz = \\
- \int q_\phi(z|x_i) \, \log \frac{p_\theta(z|x_i)}{q_\phi(z|x_i)} \, dz \\
+ \int q_\phi(z|x_i) \, \log \frac{p_\theta(z|x_i)p_\theta(x_i)}{q_\phi(z|x_i)} \, dz
\end{aligned}
\tag{1}
$$

Calculating the MLE using the first line of Equation 1 is still not tractable due to the latents being unobserved. Using Variational Inference (VI), a tractable and easy-to-sample approximate posterior distribution $q_\phi(\cdot)$ can be utilized as shown in the equation above. The second line of

Equation 1 denotes two distinct terms with the condition that $q_\phi(z|x) > 0 \iff p_\theta(z|x) > 0$. The first term denotes the KL divergence between the real and approximate posterior distributions. Minimizing this KL term between parametric posterior $q_\phi(\cdot)$ and true posterior $p_\theta(\cdot)$ would allow for sampling from $q(\cdot)$ as proxy of $p_\theta(\cdot)$, however the KL cannot be efficiently calculated due to true posterior $p_\theta(\cdot)$ not being easy to sample from. The second term is the Evidence Lower Bound (ELBo) of the likelihood which is equal to the following:

$$
\text{ELBo} = \mathbb{E}_{q_\phi(z|x_i)}[\log \, p_\theta(x|z)] - \text{KL}(q_\phi(z|x_i)||p_\theta(z))
\tag{2}
$$

The first term in the RHS of Equation 2 is the expected reconstruction of the observed data using parametric probabilistic model p_θ, under approximate density $q_\phi(\cdot)$. The seconds term encourages good prior density estimation for $q_\phi(\cdot)$, with the prior $p_\theta(\cdot)$ often being a desired distribution in practice.

A notable neural model that follows the above variational framework is Variational Auto-Encoder (VAE (Kingma & Welling, 2013)). VAE uses an encoder to parameterize the distribution $q_\phi(\cdot)$. During learning, the AEVB algorithm is used for training an encoder (or inference network) and decoder jointly together using a reparameterization trick. An alternative framework is Stochastic Variational Inference (SVI (Hoffman et al., 2013)), where the approximate posterior parameterization is done using well-known distributions as opposed to a neural model. SVI has certain appealing applications, for example SVI framework is used in Variational Auto-Decoder (VAD) for learning generative models from data with severe missingness (Zadeh et al., 2019).

2.2. Amortization Gap

In a generative modeling framework, often the decoder is considered the main component of the model. This is conventionally the component that receives samples drawn from a latent posterior distribution, and generates new data points. Using SVI (Hoffman et al., 2013) with a mean-field assumption, one can train such a model without the need for an encoder (i.e. inference) network. However, if inference is ever required during test time, such models suffer heavily due to relying on test-time gradient (or meta-gradient) descent (which is a non-parallelizable sequential operation). Using an encoder allows the process of inference to become more efficient; during test time, one can simply feed the datapoint into an encoder to get the parameters of the posterior distribution. This process is far less computationally exhaustive than gradient-based inference (since operations inside a network are usually parallelized).

However, if the inference network cannot be trained effi-

ciently - e.g. has limited capacity, or undergoes difficulties during training, or simply the nature of data is too hard for dimensionality reduction using a neural structure - then the process of learning a generative model may be suboptimal. This is called an Amortization Gap (Cremer et al., 2018), which can be somewhat mitigated using methods that require second-order gradient of the model's optimizer (Kim et al., 2018). Amortization Gap is not a theoretical weakness of inference networks (note the universal approximation theory of neural networks), but rather an empirical phenomena best describable by finite-neuron neural networks. A free-form mean-field approximate posterior inference technique such as SVI can mimic an encoder with very large capacity (due to mean-field assumption and full independence of latent parameters), and does not suffer the same gap (as shown in this paper). However, as mentioned, this comes at the cost of inference time complexity.

3. Pseudo-Encoded Stochastic Variational Inference

In this section, we outline the training process of the Pseudo-Encoded Stochastic Variational Inference (PE-SVI) framework. Training in PE-SVI is split into 3 parts: 1) Early Decoder Training, 2) Deferred Encoder Training, and 3) Pace Adjustment.

3.1. Early Decoder Training

At the first step within PE-SVI framework, a decoder is trained (without an attached encoder). Essentially, we use Stochastic Variational Inference (SVI) with mean-field assumption on latent dimensions. Assume samples $z \sim q_\phi(z|x_i)$ are drawn from a given known family of distributions (e.g. Gaussian). To generate data similar to x_i, these samples are then used as input to a decoder $\mathcal{D}_\theta(z_i)$. The reconstructed samples of this probabilistic decoder should show high resemblance such that ELBo (Equation 2) is maximized w.r.t θ, and ϕ.

$p_\theta(x|z)$ in turn can be defined as:

$$p_\theta(x|z) = \mathcal{N}\left(\mathcal{D}_\theta(z); x_i, \Lambda_i\right) \tag{3}$$

The high likelihood is therefore attributed to the low squared distance (as measure) between the output reconstruction of $\mathcal{D}_\theta(\cdot)$ and the ground-truth x_i. Λ_i is the covariance matrix of the above likelihood. The approximate posterior is not parameterized by an encoder, but rather by well-known distributions such as a Gaussian (in this paper):

$$q_\phi(z|x_i) := \mathcal{N}(z; \mu_i, \Sigma_i) \tag{4}$$

At the beginning of training, parameters of the approximate posterior $q_\phi(\cdot)$ are initialized randomly (e.g. uni-

form), same as parameters θ of the probabilistic decoder $p_\theta(\cdot)$. Within each batch of the training data, the gradient of lower-bound is calculated and the parameters of $q_\phi(\cdot)$ and $p_\theta(\cdot)$ are updated. Since there is no encoder attached to the network, backpropagation is only done to the decoders parameters (θ). In the meantime, backpropagation also happens for parameters of the approximate posterior (ϕ). Updates on the parameters of approximate posterior $q_\phi(z|x_i)$ are only done once in an epoch, when backpropagating the ELBo for x_i. Training is done until convergence w.r.t both θ and ϕ. The output of the Early Decoder Training phase is the trained approximate posteriors $q_{\phi^*}(z|x_i)$ as well as the trained decoder $\mathcal{D}_{\theta^*}(\cdot)$.

3.2. Deferred Encoder Training

After training is done, we use a neural model, also referred to as pseudo-encoder in this paper, $\mathcal{E}_\gamma(x)$ to perform a similar role as an encoder. Unlike conventional encoder-decoder architectures (in which encoder is trained end-to-end alongside the decoder) in PE-SVI, the pseudo-encoder is trained only after decoder is fully learned. The learned approximate posteriors $q_{\phi^*}(z|x_i)$ of the Early Decoder Training phase are passed to Deferred Encoder Training phase, essentially to be approximated. The objective (and a supervised one at that) is to learn a mapping between x_i and $\phi^* = \{\mu_i^*, \Sigma_i^*\}$. $\mathcal{E}_\gamma(x)$ is therefore trained in a supervised manner for this purpose, to output ϕ^* given x_i. This training can be done like any other supervision, using gradient descent approaches. After training is done $\mathcal{E}_{\gamma^*}(x)$ is used to provide a good estimate of the parameters of the true approximate posterior. For a datapoint x_i, we denote the estimates of the approximate posterior generated by $\mathcal{E}_{\gamma^*}(x)$ as $\phi^{\mathcal{E}} = \{\mu_i^{\mathcal{E}}, \Sigma_i^{\mathcal{E}}\}$.

3.3. Pace Adjustment

For ith input x_i, the parameters of the approximate posterior $q_{\phi^{\mathcal{E}}}(z|x_i)$ are first obtained using the pseudo-encoder $\mathcal{E}_{\gamma^*}(x_i)$. Subsequently, these parameters can be refined using SVI to achieve the final posteriors $q_{\phi^*}(z|x_i)$. This by itself reduces the number of SVI steps required to maximize the ELBo to a significant amount (naturally due to approximation of the $\phi^* = \{\mu_i^*, \Sigma_i^*\}$ using $\phi^{\mathcal{E}} = \{\mu_i^{\mathcal{E}}, \Sigma_i^{\mathcal{E}}\}$, also shown in experiments in this paper). However, during Pace Adjustment phase, one can switch to SVI step[1] sizes that are most suited for convergence, given the initial estimates of approximate posterior parameters $\phi^{\mathcal{E}}$. Therefore, higher learning rates, which are often detrimental if approximate posterior is initialized randomly, can be used to maximize ELBo w.r.t ϕ (initialized with $\phi^{\mathcal{E}}$). Thus, a further reduction in number of steps can be made by simply taking

---

[1]In this paper, we use Adam (Kingma & Ba, 2014) as the optimizer for approximate posterior parameters.

larger steps. One can simply treat the adjusted learning rate as a hyperparameter, and pick the one with fastest and most accurate convergence to ϕ^*. Any hyperpatameter optimization method (e.g. Bayesian hyperparameter optimization approaches (Snoek et al., 2012)) may be used for more accurate localization of a suitable pace. For the sake of this paper, we simply suffice to treating the adjusted learning rate as a hyperparameter found using random (yet sensible) grid search.

4. Experiments

In this section, we discuss the details of the experiments for this paper. We first start by discussing the studied datasets, followed by methodology and hyperparameter space search.

4.1. Datasets

We use the following set of datasets in our experiments:

Synthetic Data: As the first dataset in our experiment, we study a case of synthetic data where we control the distributional properties of the data. In the generation process, we first acquire a set of independent dimensions randomly sampled from 5 univariate distributions with uniform random parameters: {`Normal`, `Uniform`, `Beta`, `Logistic`, `Gumbel`}. Often in realistic scenarios there are inter-dependencies among the dimensions. Hence we proceed to generate interdependent dimensions by picking random subsets of the independent components and combining them using random operations such as weighted multiplication, affine addition, and activation. Using this method, we generate a dataset containing $N = 50,000$ datapoints with ground-truth dimension $d = 300$.

CMU-MOSI Dataset: CMU Multimodal Opinion Sentiment Intensity (CMU-MOSI) is a dataset of multimodal language specifically focused on multimodal sentiment analysis (Zadeh et al., 2016). It is among the most well-studied multimodal language datasets in NLP community. Multimodal sentiment analysis extends conventional language-based sentiment analysis to a multimodal setup where both verbal and non-verbal signals contribute to the expression of sentiment. CMU-MOSI contains 2199 video segments taken from 93 Youtube movie review videos. The train, validation and test folds of the CMU-MOSI contain 1248, 229 and 686 segments respectively (Chen et al., 2017). We use expected multimodal context for each sentence, similar to unordered compositional approaches in NLP (Iyyer et al., 2015).

300-W: (Sagonas et al., 2013a;b) is a meta-dataset of four different facial landmark datasets: Annotated Faces in the Wild (AFW) (Zhu & Ramanan, 2012), iBUG (Sagonas et al., 2013c), and LFPW + He-

Model \ $|z|$	16	32	64	128
CMU-MOSI				
VAE	0.7176	0.5871	0.4681	0.2623
SVI	0.0470	0.0010	0.0006	0.0003
PE-SVI-0	0.0516	0.0064	0.0090	0.0063
PE-SVI-25	0.0482	0.0010	0.0007	0.0003
300-W				
VAE	0.2349	0.2123	0.1450	0.0922
SVI	0.0012	0.0009	0.0006	0.0004
PE-SVI-0	0.0798	0.0775	0.0697	0.0592
PE-SVI-25	0.0011	0.0008	0.0007	0.0002
Synthetic				
VAE	78.6053	73.9616	66.1229	60.1511
SVI	0.0331	0.0117	0.0021	0.0005
PE-SVI-0	0.9706	0.5561	0.5282	0.4197
PE-SVI-25	0.0348	0.0119	0.0039	0.0027
SST				
VAE	0.4860	0.4162	0.3801	0.3233
SVI	0.1411	0.1228	0.0895	0.0506
PE-SVI-0	0.3781	0.3605	0.3559	0.3499
PE-SVI-25	0.1417	0.1229	0.0887	0.0517

Table 1. The results of experiments on Arch1. Refer to Section 5 for discussion and analysis.

len (Belhumeur et al., 2011; Le et al., 2012) datasets. We used the full iBUG dataset and the test partitions of LFPW and HELEN. This led to 135, 224, and 330 images for testing respectively. They all contain uncontrolled images of faces in the wild: in indoor-outdoor environments, under varying illuminations, in presence of occlusions, under different poses, and from different quality cameras. For the purpose of statistical shape modeling, only the landmarks are used.

SST: The Stanford Sentiment Treebank (SST) is a dataset of movie review excerpts from Rotten Tomatoes website (Socher et al., 2013). The dataset is annotated for both root and intermediate nodes of parsed sentences. We only use the root nodes in our experiments. Similar to CMU-MOSI, we use an unordered compositional approach for the input sentence embeddings.

4.2. Methodology

For all the datasets, we study the following feed-forward encoder-decoder or decoder-only architectures. For all the architectures, $|z|$ is the dimensionality of the latent space. The encoder has the same architecture as the decoder, only inverted. The following decoder architectures

Model \ $|z|$	16	32	64	128
CMU-MOSI				
VAE	0.5778	0.3644	0.2767	0.2257
SVI	0.0642	0.0170	0.0020	0.0015
PE-SVI-0	0.0686	0.0214	0.0068	0.0060
PE-SVI-25	0.0644	0.0171	0.0020	0.0019
300-W				
VAE	0.1711	0.1279	0.1090	0.0489
SVI	0.0022	0.0014	0.0012	0.0010
PE-SVI-0	0.0698	0.0692	0.0669	0.0614
PE-SVI-25	0.0047	0.0042	0.0031	0.0020
Synthetic				
VAE	47.6520	29.7762	23.5845	17.8166
SVI	0.0940	0.0491	0.0172	0.0155
PE-SVI-0	0.5445	0.5283	0.5242	0.4968
PE-SVI-25	0.0730	0.0530	0.0292	0.0156
SST				
VAE	0.4552	0.3994	0.3040	0.2576
SVI	0.1718	0.1434	0.1302	0.0808
PE-SVI-0	0.3951	0.3624	0.3268	0.2417
PE-SVI-25	0.1728	0.1444	0.1273	0.0804

Table 2. The results of experiments on Arch2. Refer to Section 5 for discussion and analysis.

are used in this paper: [Arch1] $\mathcal{D}_\theta^{A1}(z) : z \mapsto x$, [Arch2] $\mathcal{D}_\theta^{A2}(z) : z \mapsto \mathbb{R}^{min(z \times 2, 128)} \mapsto x$, [Arch3] $\mathcal{D}_\theta^{A3}(z) : z \mapsto \mathbb{R}^{min(z \times 2, 128)} \mapsto \mathbb{R}^{min(z \times 2, 128)} \mapsto x$. All the models are ReLU activated.

The following models are studied in this paper:

VAE: Variational Auto-Encoder uses and encoder to perform posterior approximation and a decoder to reconstruct a given input. Encoder and decoder are trained together end to end. The amortization gap essentially may happen during training (Cremer et al., 2018).

SVI: We use Stochastic Variational Inference directly on free-form latent parameters. We make a mean-field assumption for amortizing the posterior approximation. The free parameters of the latent space are essentially the parameters of a Gaussian distribution.

PE-SVI-0: Essentially, this is the proposed model in this paper without the adjustment steps in Section 3.3. The latent inference is simply done using the trained encoder in Section 3.2.

PE-SVI-25: This is essentially PE-SVI-0, with 25 steps with adjusted learning rate as discussed in Section 3.3.

For all the models, we assume no particular prior distribution for latent space, therefore, in this paper we are only concerned with expected likelihood under the approximate posterior distribution (first term of ELBo in Equation 2). This essentially compares the models for their reconstruction power. Note, we do not argue that a good generative model has more properties than just good reconstruction; however, good reconstruction is required for good generative modeling. In theory, the second term in ELBo has no direct dependency on the reconstruction as it simply forces the latent space to follow a particular distribution. This term is the same for both SVI and VAE, and therefore, both models can be adapted to follow a particular desired latent space distribution. To compare the reconstruction performance of both models, we directly maximize the expected log-likelihood reconstruction term within ELBo, and report the negative of its value.

4.3. Hyperparameter Space Search

The VAE models in this paper are trained using Adam with learning rates $\{1, 5, 8\} \times 10e - \{2, 3, 4, 5\}$ for a total of 3000 epochs. SVI and PE-SVI models are trained using $10e - \{2, 3\}$ for model parameters and $10e - \{1, 2, 3\}$ for latent parameters (model and latent learning rates are independent). The best models are picked based on the performance on validation-set, and directly applied to the test-set of each dataset (random 10% held out for validation and test). The Reduced Adaptation Steps are a total of 25 epochs and the learning rates of $\{1, 5\} \times 10e - \{0, 1, 2\}$. The hyperparameter space is searched with $12\times$ Tesla V100 gpus.

5. Results and Discussion

The results of experiments over all datasets, baselines and architectures are reported in Tables 1, 2, 3 respectively for Arch1,2,3. We report the observations from these tables as follows:

Performance Comparison (VAE, SVI): Firstly, we report whether or not a gap exists between SVI and VAE performance. Tables 1, 2, 3 demonstrates superior performance for SVI over VAE, by a rather large margin in certain cases. This gap signals a performance suboptimality for VAE model, also observed in previous works (Cremer et al., 2018; Hjelm et al., 2016).

Performance Comparison (VAE, PE-SVI-0): Surprisingly, we observe that in majority of our experiments, PE-SVI-0 performs better than VAE. Both models use an identical encoder architecture to perform approximate posterior inference. However, the decoder training is different across the models. We suspect that the lack of a performance gap when training using SVI (in Early Decoder Training phase),

Model \ $|z|$	16	32	64	128
CMU-MOSI				
VAE	0.4900	0.3623	0.2180	0.2771
SVI	0.0980	0.0835	0.0032	0.0026
PE-SVI-0	0.1062	0.0870	0.0061	0.0065
PE-SVI-25	0.0987	0.0837	0.0033	0.0027
300-W				
VAE	0.1466	0.1142	0.0742	0.0351
SVI	0.0021	0.0013	0.0011	0.0009
PE-SVI-0	0.0489	0.0492	0.0475	0.0146
PE-SVI-25	0.0046	0.0041	0.0030	0.0020
Synthetic				
VAE	35.4315	29.5171	25.6991	25.4315
SVI	0.1209	0.0937	0.0354	0.0185
PE-SVI-0	0.6167	0.5527	0.5262	0.4420
PE-SVI-25	0.1181	0.0922	0.0310	0.0227
SST				
VAE	0.3618	0.2966	0.2008	0.1611
SVI	0.2140	0.1871	0.1462	0.1010
PE-SVI-0	0.5243	0.4633	0.4871	0.5589
PE-SVI-25	0.3149	0.1872	0.1504	0.1098

Table 3. The results of experiments on Arch3. Refer to Section 5 for discussion and analysis.

allows subsequent training of the encoder (in Deferred Encoder Training phase) to be more successful; as compared to training with both which can essentially lead to suboptimal performance for both encoder and decoder. It should be noted that the ultimate purpose of PE-SVI is to reduce the steps required for SVI inference, and this comparison was made as byproduct of our experiments.

Performance Comparison (SVI, PE-SVI-25): PE-SVI-25, which performs 25 adjusted steps (see Section 3.3) after PE-SVI-0, is able to closely approximate the performance of the SVI model. For SVI, the number of required steps for inference convergence is usually higher than 1000 across all our datasets. For example, convergence steps for SST Arch1 (Table 1) with $|z| = 128$ is 2381 with learning rate of 0.001 (non-convergent with 0.01), while PE-SVI-25 reaches the same performance in 12 steps (and plateaus afterwards) with learning rate of 0.1. Thus higher learning rate (different than used for random initialization) can successfuly be used with PE-SVI, after Pace Adjustment.

Performance Comparison (SVI, PE-SVI-0): The comparison between SVI and PE-SVI-0 suggests the latent space learned by SVI is complex, and not perfectly reconstructable using an encoder, which naturally has limited

inference capacity. Such a suboptimality naturally takes a toll at the training process (Hjelm et al., 2016), as also observed from comparison between SVI and VAE.

Performance Comparison (SVI across Arch1,2,3): Surprisingly, depth of the decoder seems to negatively impact the performance of SVI. This demonstrates that in many cases, a small decoder may be enough to learn a generative model with good reconstruction. However, this depth positively impacts PE-SVI-0 and VAE, signaling that more powerful encoders are better capable at approximating the latent space learned by SVI. Regardless, PE-SVI-25 follows the performance of SVI. We did not observe any pattern in higher or lower number of epochs required for convergence of PE-SVI-25 across different depths.

Performance Comparison ($|z|$): Unanimously across all models, architectures and datasets, increasing the dimensionality of the latent space improves the performance.

6. Conclusion

In this paper, we presented a new approach called Pseudo-Encoded Stochastic Variational Inference (PE-SVI), to reduce the inference complexity of SVI during test time. Our approach relies on finding a suitable initial start point for gradient operations, which naturally reduces the required SVI steps. Furthermore, this suitable start point allows for taking larger SVI step sizes during test-time inference (compared to random initialization) which further reduces the required SVI steps. Essentially, we learn a parametric model to output this start point. In our experiments, PE-SVI achieves similar performance to SVI, however with a fraction of required inference steps. Furthermore, we observe that the initial PE-SVI start point (without any SVI steps) shows better performance than jointly training a decoder with an inference model (e.g. VAE).

References

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

Chen, M., Wang, S., Liang, P. P., Baltrušaitis, T., Zadeh, A., and Morency, L.-P. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 163–171. ACM, 2017.

Cremer, C., Li, X., and Duvenaud, D. Inference suboptimality in variational autoencoders. *arXiv preprint arXiv:1801.03558*, 2018.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. The

helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

Hjelm, D., Salakhutdinov, R. R., Cho, K., Jojic, N., Calhoun, V., and Chung, J. Iterative refinement of the approximate posterior for directed belief networks. In *Advances in Neural Information Processing Systems*, pp. 4691–4699, 2016.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pp. 1681–1691, 2015.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kim, Y., Wiseman, S., Miller, A. C., Sontag, D., and Rush, A. M. Semi-amortized variational autoencoders. *arXiv preprint arXiv:1802.02550*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Le, V., Brandt, J., Lin, Z., Bourdev, L., and Huang, T. S. Interactive facial feature localization. In *Computer Vision– ECCV 2012*, pp. 679–692. Springer, 2012.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 397–403, 2013a.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W), Workshop on Analysis and Modeling of Faces and Gestures*, 2013b.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV*, 2013c.

Snoek, J., Larochelle, H., and Adams, R. P. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pp. 2951–2959, 2012.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder variational autoencoders. In *Advances in neural information processing systems*, pp. 3738–3746, 2016.

Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

Zadeh, A., Lim, Y.-C., Liang, P. P., and Morency, L.-P. Variational auto-decoder: Neural generative modeling from partial data. *arXiv preprint arXiv:1903.00840*, 2019.

Zhu, X. and Ramanan, D. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2879–2886. IEEE, 2012.