Robust Variational Inference

Michael Figurnov Kirill Struminsky Dmitry Vetrov
michael@figurnov.ru k.struminsky@gmail.com vetrovd@yandex.ru

Motivation

- Real-world datasets often include outliers and noisy objects. Cleaning the data might be impractical
- Suppose that our probabilistic model can only deal with the clean objects
- We develop a scalable **robust** inference procedure that ignores the objects which cannot be explained by the data model (objects with low evidence)

Robust model evidence

- $p(x_i|\theta)$ is evidence for a data point x_i for a model with parameters θ
- The robust evidence is obtained by adding a regularization coefficient $\varepsilon > 0$ to the evidence:

$$\sum_{i=1}^{N} \log p(x_i|\theta) \to \sum_{i=1}^{N} \log(\varepsilon + p(x_i|\theta)) \quad (1)$$

  to define the robust model evidence
- The robust model evidence penalizes the model for small $p(x_i|\theta)$ less. If $p(x_i|\theta) \ll \varepsilon$, the evidence can take arbitrarily small values, while the robust evidence is bounded from below $\log(\varepsilon + p(x_i|\theta)) > \log \varepsilon$.
- The choice of ε is important. Intuitively, the higher the ε, the more training objects are ignored

Robust evidence lower bound $\mathcal{L}_\varepsilon$

- Consider a model with local latent variables z (*e.g.*, variational autoencoder)

$$p(X, Z|\theta) = \prod_{i=1}^{N} p(x_i, z_i|\theta) \quad (2)$$

- Standard evidence lower bound $\mathcal{L}$:

$$\mathcal{L}(X, \theta, \phi) = \sum_{i=1}^{N} \mathbb{E}_{q(z_i|x_i,\phi)} \log \frac{p(x_i, z_i|\theta)}{q(z_i|x_i, \phi)} \le \sum_{i=1}^{N} \log p(x_i|\theta) \quad (3)$$

  for any variational distribution $q(z_i|x_i, \phi)$
- Robust evidence lower bound $\mathcal{L}_\varepsilon$:

$$\mathcal{L}_\varepsilon(X, \theta, \phi) = \sum_{i=1}^{N} \mathbb{E}_{q(z_i|x_i,\phi)} \log \left[\varepsilon + \frac{p(x_i, z_i|\theta)}{q(z_i|x_i, \phi)}\right] \le \sum_{i=1}^{N} \log\left[\varepsilon + p(x_i|\theta)\right] \quad (4)$$

  Proof:

$$\log\left[\varepsilon + p(x_i|\theta)\right] = \log\left[\mathbb{E}_{q(z_i|x_i,\phi)}\left(\varepsilon + \frac{p(x_i, z_i|\theta)}{q(z_i|x_i, \phi)}\right)\right]$$

$$\{\text{Jensen's inequality}\} \ge \mathbb{E}_{q(z_i|x_i,\phi)} \log\left[\varepsilon + \frac{p(x_i, z_i|\theta)}{q(z_i|x_i, \phi)}\right]$$

- Both $\mathcal{L}$ and $\mathcal{L}_\varepsilon$ can be optimized with stochastic gradient ascent by using the reparametrization trick

Analysis of the robust evidence lower bound $\mathcal{L}_\varepsilon$

- For a fixed x_i, z_i, the gradients of $\mathcal{L}$ and the robust version $\mathcal{L}_\varepsilon$ have the same direction but different magnitudes:

$$\nabla \log\left[\varepsilon + \frac{p(x_i, z_i|\theta)}{q(z_i|x_i, \phi)}\right] = \left(\frac{\frac{p(x_i,z_i|\theta)}{q(z_i|x_i,\phi)}}{\varepsilon + \frac{p(x_i,z_i|\theta)}{q(z_i|x_i,\phi)}}\right) \nabla \log\left[\frac{p(x_i, z_i|\theta)}{q(z_i|x_i, \phi)}\right] \quad (5)$$

- **The unlikely objects contribute less to the gradients**
  - When $\frac{p(x_i,z_i|\theta)}{q(z_i|x_i,\phi)} \ll \varepsilon$, the scalar factor is close to zero.
  - When $\frac{p(x_i,z_i|\theta)}{q(z_i|x_i,\phi)} \gg \varepsilon$, it is close to one.

Choosing the robustness parameter ε

- The choice of ε depends on the current evidence of the dataset which changes during training
- We choose the following form of ε:

$$\varepsilon = \alpha \exp\left(\frac{\mathcal{L}(X, \theta, \phi)}{|X|}\right), \ \alpha > 0 \quad (6)$$

- In practice, we estimate the mean evidence lower bound using exponential moving average with $\gamma = 0.99$. We update ε after each gradient step

Noisy data experiment

- MNIST and OMNIGLOT datasets with stochastic binarization (pixels are Bernoulli random variables with $p =$ intensity) as in (Burda et al., 2016)
- Noise object: intensity of all pixels is the mean intensity of the training set
- Model: variational auto-encoder (VAE) with 50 Gaussian latent variables
- Robust VAE is trained with $\mathcal{L}_\varepsilon$, VAE with $\mathcal{L}$
- Robust VAE outperforms VAE for a wide range of α

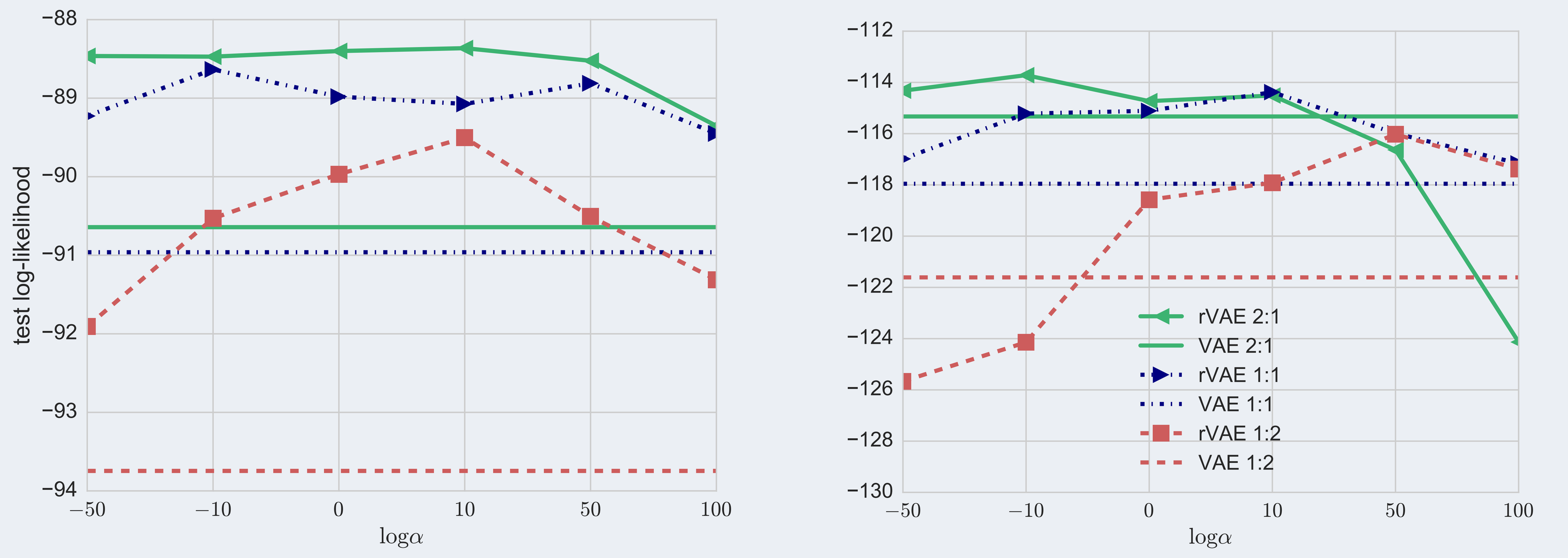

Figure: **Left**: MNIST, **right**: OMNIGLOT. Log-likelihood estimates of robust (**rVAE**) and non-robust models (**VAE**) of the clean **test set**. Models were trained on synthetically corrupted datasets, labels specify (data : noise) ratio in the experiments.

- Robust VAE describes *noise* significantly worse than VAE

	Robust VAE						VAE
$\log\alpha$	-50	-10	0	10	50	100	-
MNIST	-307.23	-308.33	-312.98	-308.15	-395.96	**-441.67**	-304.76
OMNIGLOT	-224.80	-227.94	-229.85	-241.21	-359.94	**-397.94**	-224.75

Table: Log-likelihood estimates of **synthetic noise**. The ratio (data : noise) is fixed to (1:2)

Pure data experiment

- We trained the VAE and Robust VAE models from the previous experiment on the uncorrupted datasets. Robust VAE slightly outperforms VAE in this setting for low α, suggesting a regularization effect

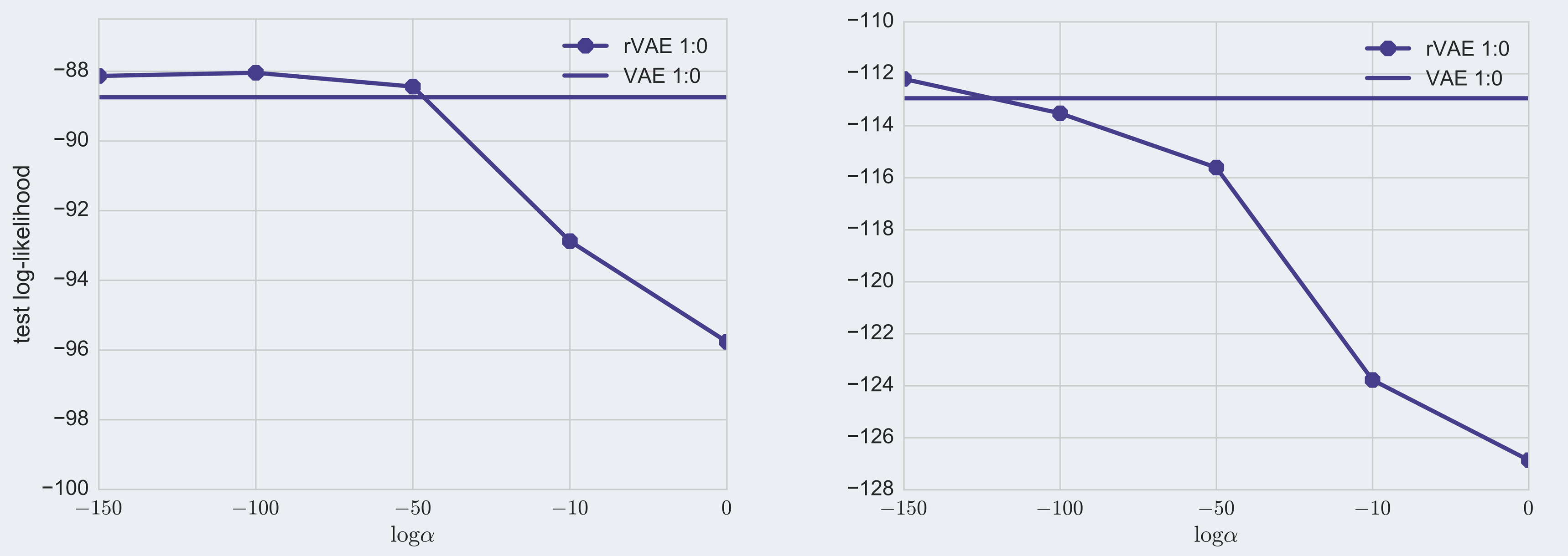

Figure: **Left**: MNIST, **right**: OMNIGLOT. Log-likelihood estimates of robust (**rVAE**) and non-robust (**VAE**) models on the **test set**

Future work

- Design an inference procedure for ε
- Compare to Wang et al. (2016)
- Evaluate on other datasets

References

Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," *ICLR*, 2016.
Y. Wang, A. Kucukelbir, and D. M. Blei, "Reweighted data for robust probabilistic models," *arXiv preprint arXiv:1606.03860*, 2016.