1 Derivation of likelihood

We are given $p(y|x,z) = \sigma(yz^T x)$, where $\sigma(x) = 1/(1+\exp(-x))$.

$$\begin{aligned}\log p(y|x,z) &= \log \frac{1}{1+\exp(-yz^T x)} \\ &= \log 1 - \log(1+\exp(-yz^T x)) \\ &= 0 - \log(\exp(0)+\exp(-yz^T x)) \\ &= -lae(0,-yz^T x) \ \ \text{,where } \ lae(s,t) = \log(\exp(s)+\exp(t))\end{aligned} \tag{1}$$

2 Derivation of gradient

$$\begin{aligned}\frac{\partial}{\partial t} lae(s,t) &= \frac{\partial}{\partial t}\log(\exp(s)+\exp(t)) \\ &= \frac{1}{\exp(s)+\exp(t)}\frac{\partial}{\partial t}(\exp(s)+\exp(t)) \\ &= \frac{\exp(t)}{\exp(s)+\exp(t)} \\ &= \frac{1}{1+\exp(s-t)} = \sigma(t-s)\end{aligned} \tag{2}$$

$$\begin{aligned}\nabla_z \log p(y|x,z) &= \nabla_z - lae(0,-yz^T x) \\ &= -\nabla_z lae(0,-yz^T x) \\ &= -\sigma(-yz^T x - 0)\nabla_z(-yz^T x) \\ &= -\sigma(-yz^T x) - yx \\ &= \sigma(-yz^T x)yx\end{aligned} \tag{3}$$

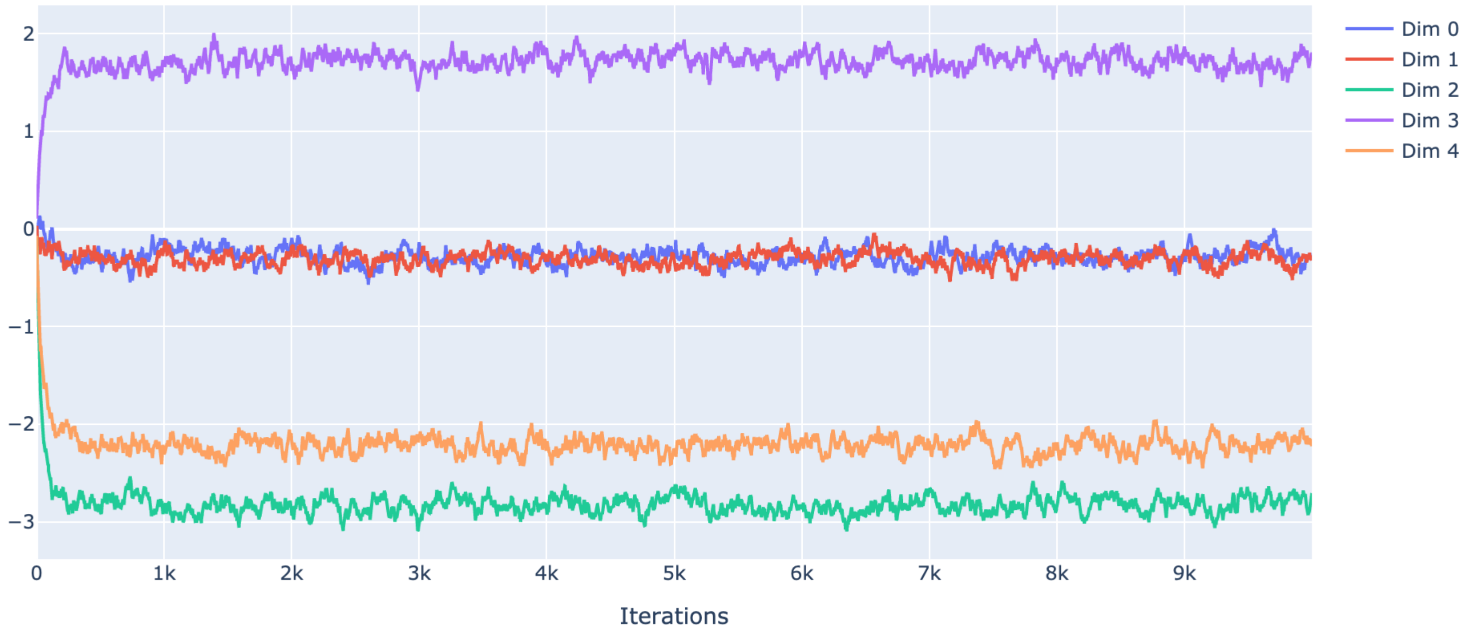
3 Pseudocode for Stochastic gradient variational inference

We use $\mathcal{N}(w, 0.5I)$ as our variational distribution $q_w(z)$. We can approximately compute $\int q_w(z) \log p(z, Data) \mathrm{d}z$ by taking samples from $q_w(z)$ and computing $\log p(z, Data)$. Since x is observed, we avoid modeling it. Further we assume the prior to be a standard unit normal.

```
Algorithm 1: Estimating Mean Evidence
  Input: w
  $z \backsim q_w(z) \backsim \mathcal{N}(w, 0.5I)$ ;
  Compute $\log p(z, Data)$ as ;
  $= \log \left( p(z) \prod_{i=1}^{N} \log p(y^{(i)}|x^{(i)}, z) \right) = \log \exp(\frac{-1}{2}\|z\|^2) + \sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)}, z)$ ;
  $evidence = \frac{-1}{2}\|z\|^2 + \sum_{i=1}^{N} lae(0, -yz^T x^{(i)})$ ;
  return evidence
```

```
Algorithm 2: Gradient of Mean Evidence
  Input: w
  $z \backsim \mathcal{N}(w, 0.5I)$ ;
  /* Using reparameterization trick, ignoring entropy due to fixed covariance */
  $\nabla_w \mathbb{ELBO} = \nabla_w \frac{-1}{2}\|T_w(\epsilon)\|^2 + \nabla_w \sum_{i=1}^{N} lae(0, -yT_w(\epsilon)^T x^{(i)})$ ;
  /* Using $T_w(\epsilon) = z$ */
  $\nabla_w \mathbb{ELBO} = -z + \sum_{i=1}^{N} \sigma(-y^{(i)} z^T x^{(i)}) y^{(i)} x^{(i)}$ ;
  return $\nabla_w \mathbb{ELBO}$
```

4 Implement SGVI



5 Direct predictive accuracy for SGVI

5.1 Prediction using samples

For a new test input x, we sample $z_1, z_2, ...z_{t_{max}}$ from the posterior $p(z|Data)$. We compute the mean of each prediction, which is the sigmoid of the dot product of the sample and test input.

$$p(y = 1) = \frac{1}{t_{max}} \sum_{i=1}^{t_{max}} \sigma(x^T z_i) \tag{4}$$

5.2 Error rates

		Repetitions				
Iterations	10	0.149	0.205	0.155	0.153	0.159
	100	0.147	0.144	0.146	0.15	0.144
	1000	0.143	0.147	0.145	0.139	0.14
	10000	0.145	0.144	0.145	0.144	0.143

Table 1 – Error rates for increasing iterations across 5 repetitions

6 Step Sizes and Run Lengths for SGVI Dynamics

We experiment with different step sizes and iterations. To measure performance, we evaluate on the test set at specified intervals. We plot the error rates subsequently.

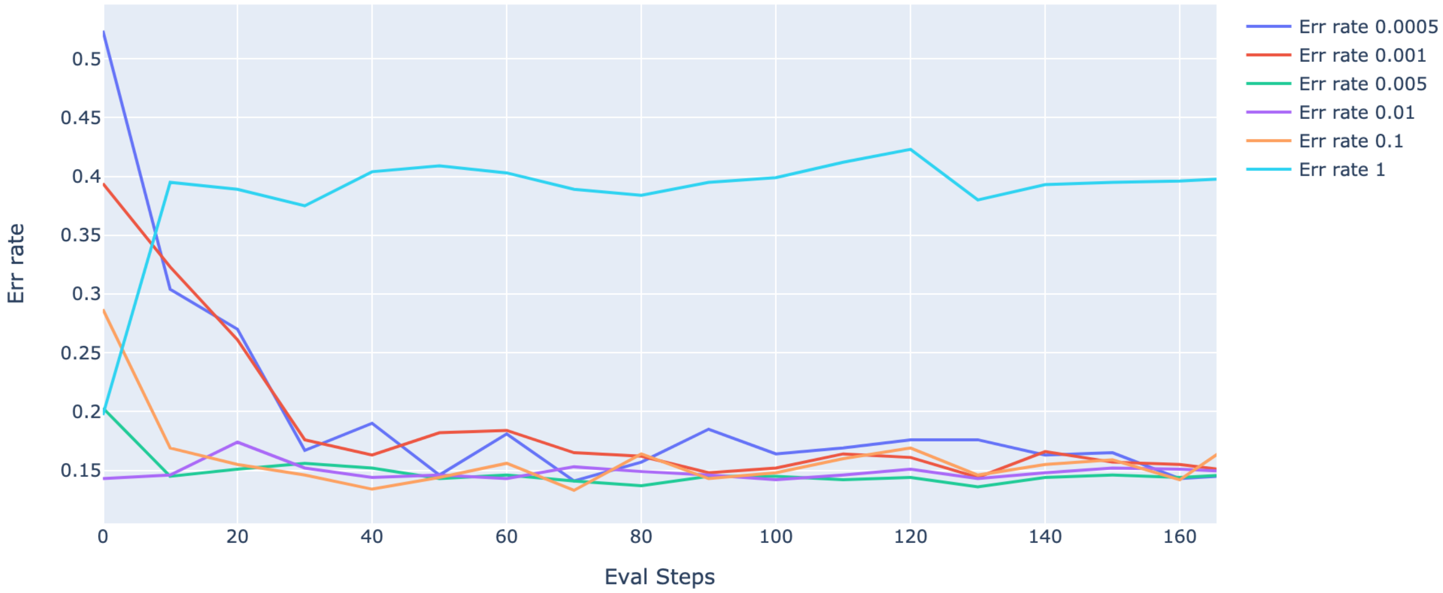


Figure 1 – Average performance as measured by error rate for different step sizes. Number of iterations fixed to 1000, evaluated every 20 steps

We see that the error rates converges after approximately 1000 iterations. As expected a larger step size, leads to faster convergence. However, setting a very large step-size = 1, causes divergence, as seen from the figure.

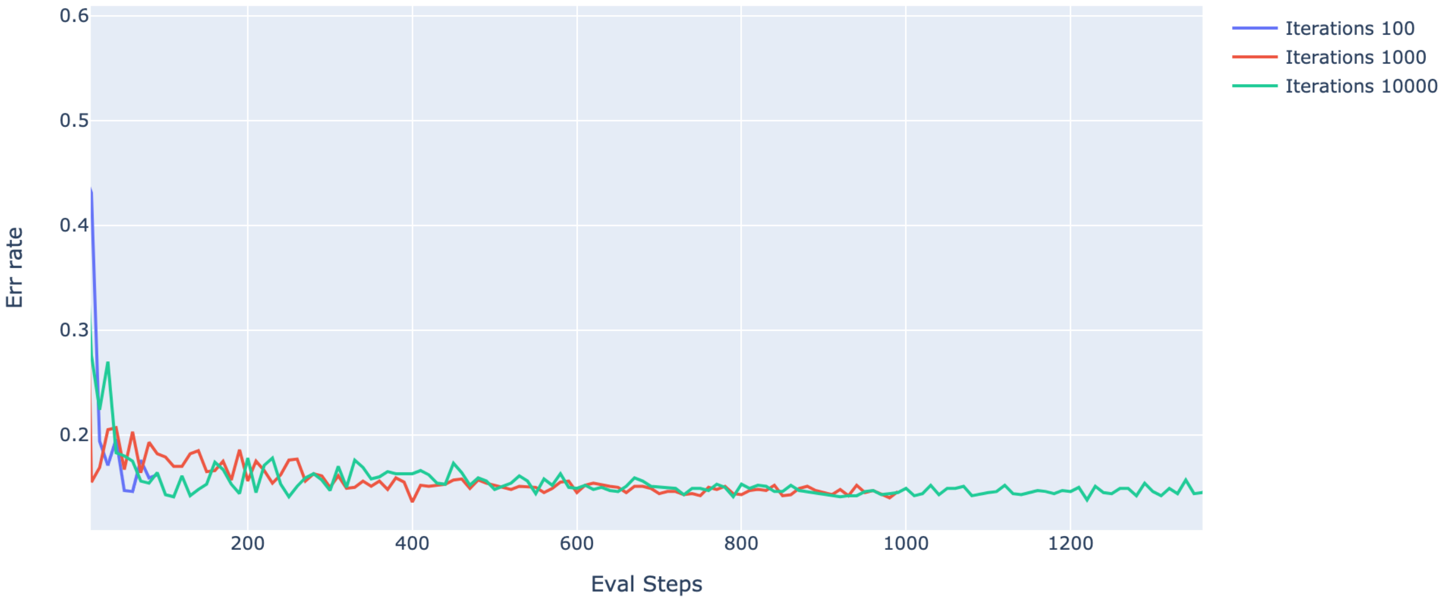


Figure 2 – Checking convergence acorss different run lenghts