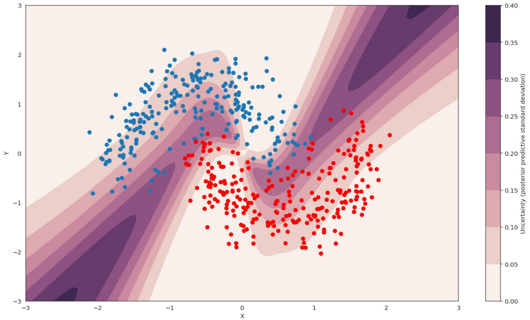
Variational deep learning

Dr. Luca Ambrogioni

Part I: Introduction to variational inference

Dr. Luca Ambrogioni

Back to Bayesian neural networks



Forward pass:

$$y_j = W_2 f(W_1 x_j + b_1) + b_2$$

Likelihood (binary classification):

$$p(D|W_1, W_2, b_1, b_2) = \prod_{j=1}^{N} \rho(x_j)^{y_j} (1 - \rho(x_j))^{1-y_j}$$

Prior:

$$p(W_1, W_2, b_1, b_2) = \mathcal{N}(W_1|0, I)\mathcal{N}(W_2|0, I)\mathcal{N}(b_1|0, 10I)\mathcal{N}(b_2|0, 10I)$$

Learning as posterior inference

Posterior over the network parameters (learning target)

$$p(W_1, W_2, b_1, b_2 | D) \propto$$
$$p(D | W_1, W_2, b_1, b_2) p(W_1) p(W_2) p(b_1) p(b_2)$$

Likelihood (exp of the loss)

Prior over the parameters (regularization)

Learning as posterior inference

Collected parameters

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{p(D)}$$

Data

Marginal likelihood (Normalization factor)

The intractability of Bayesian inference

Intractable integral (≈ weighted sum) over high-dimensional space

$$p(D) = \int_{\text{parameter space}} p(D|\Theta)p(\Theta)\mathrm{d}\Theta$$

Bayesian inference as optimization problem

Learnable distribution (variational approximation)

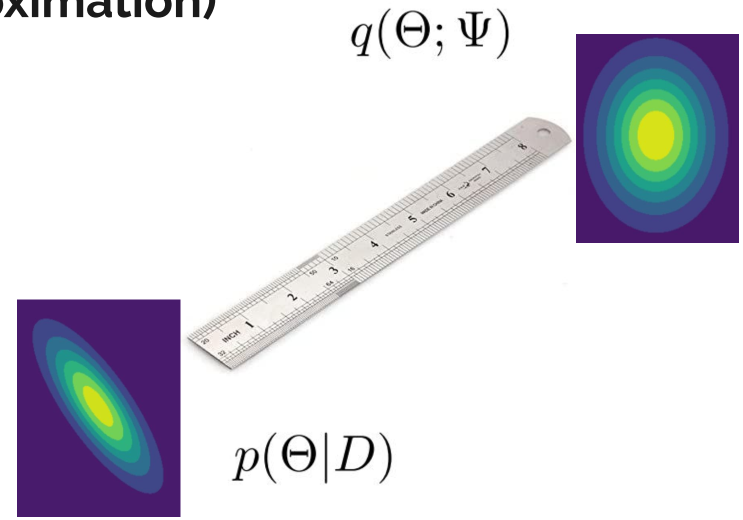
$$q(\Theta; \Psi)$$

Target distribution (true posterior)

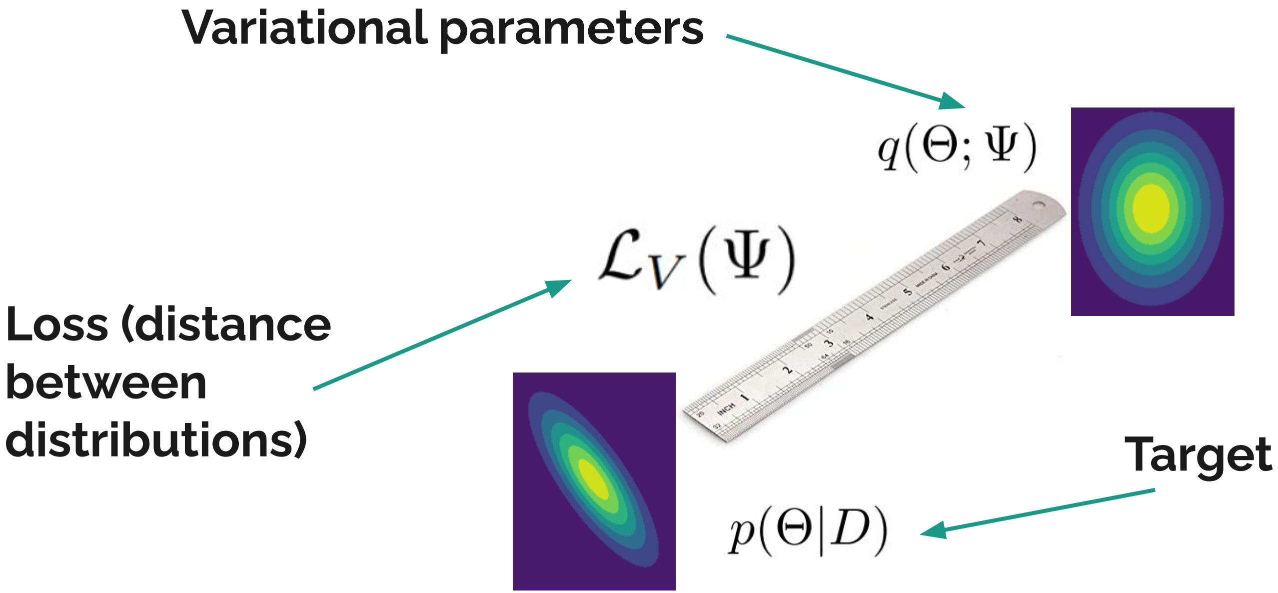
$$p(\Theta | D)$$

Distributional loss function(al)

$$\mathcal{L}_V(\Psi) = D\left(p(\Theta|D), q(\Theta; \Psi)\right)$$

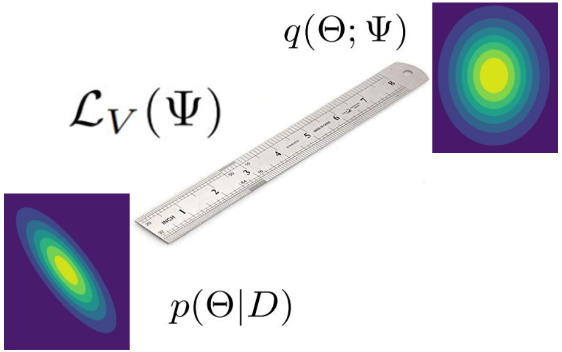


Bayesian inference as optimization problem



Requirements of the loss functional

1. D(p,q) = 0 if and only if p = q
2. D(p, q) > 0 if p ≠ q

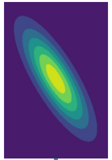


Bayesian inference by gradient descent

Gradient of the distributional loss

$$\Psi_{n+1} = \Psi_n - \eta \nabla \mathcal{L}_V(\Psi_n)$$

Learning rate

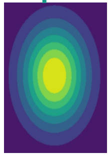


$q(\Theta|\Psi_N)$

Trained approximation

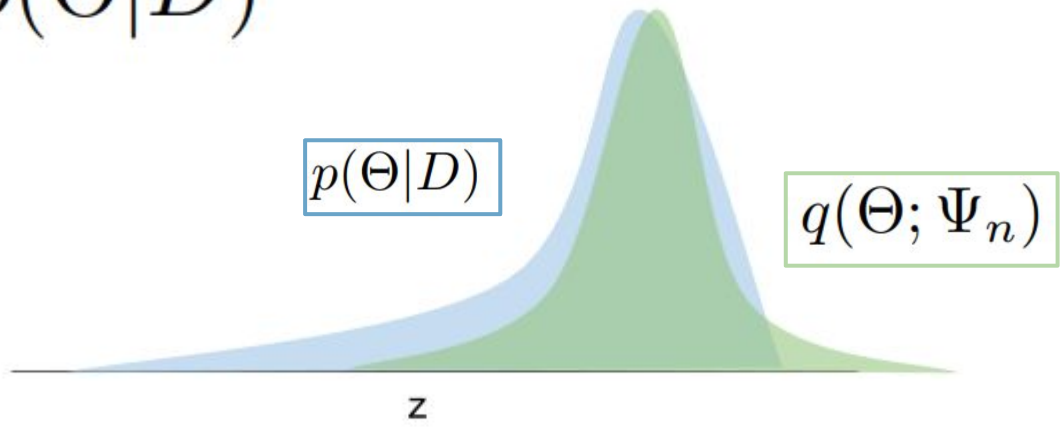
$q(\Theta|\Psi_0)$

Starting point



Variational approximation

$$q(\Theta; \Psi_n) \approx p(\Theta | D)$$

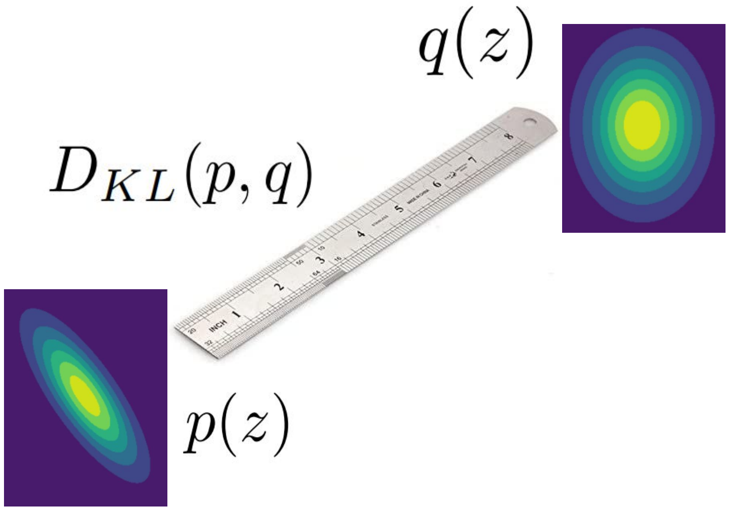


Interlude: The KL divergence

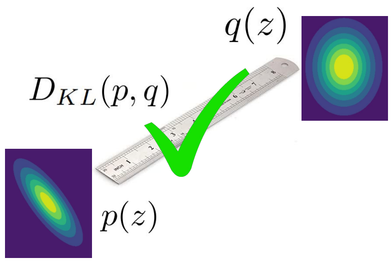
Dr. Luca Ambrogioni

The KL divergence: a measuring rule between probability distributions

$$D_{\mathrm{KL}}(p, q) = \mathbb{E}_{z \sim p}\left[\log \frac{p(z)}{q(z)}\right]$$



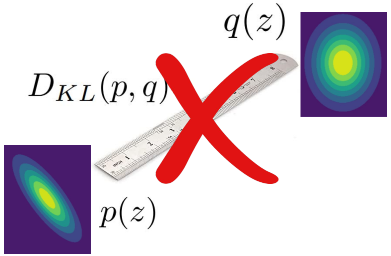
The KL divergence: basic properties



$$D_{\mathrm{KL}}(p,q) \geq 0, \quad \text{for all distributions } p(z), q(z)$$

$$D_{\mathrm{KL}}(p,q) = 0, \quad \text{if and only if } p(z) \text{ is identical to } q(z)$$

The KL divergence: not symmetric!



$$D_{\mathrm{KL}}(p,q) \neq D_{\mathrm{KL}}(q,p)$$

A loss on the space of probability distributions!

$$q(z; W)$$

Parameterized distribution

$$\mathcal{L}(W) = -D_{\mathrm{KL}}(p_{\mathrm{target}}(z), q(z; W))$$

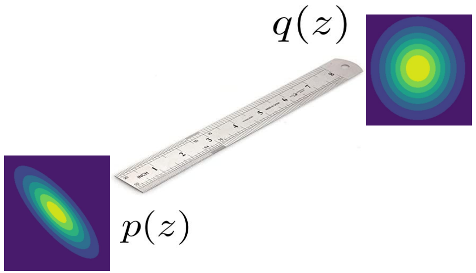
Target distribution

Part II: Reversed KL variational inference

Dr. Luca Ambrogioni

Using the (reversed) KL divergence as functional loss

$$\mathcal{L}_V(\Psi) = D_{\mathrm{KL}}(q, p) = \mathbb{E}_{q(\Theta;\Psi)}\left[\log \frac{q(\Theta;\Psi)}{p(\Theta \mid D)}\right]$$



Expectation over the variational approximation

Is the KL loss tractable??

It seems that in order to evaluate the loss we need to know the intractable true posterior!

$$\mathcal{L}_V(\Psi) = D_{\mathrm{KL}}(q, p) = \mathbb{E}_{q(\Theta;\Psi)}\left[\log \frac{q(\Theta;\Psi)}{p(\Theta \mid D)}\right]$$

If so, the whole approach is useless as approximating the posterior was our original problem!

Is the KL loss tractable?? Yes! (up to a constant!)

$$\boxed{\log \frac{a}{\left(\frac{b}{c}\right)} = \log \frac{ac}{b} = \log \frac{a}{b} + \log c}$$

Tractable term that depends on Psi

$$\mathbb{E}_{q(\Theta;\Psi)}\left[\log \frac{q(\Theta;\Psi)}{\frac{p(D|\Theta)p(\Theta)}{p(D)}}\right] = \mathbb{E}_{q(\Theta;\Psi)}\left[\log \frac{q(\Theta;\Psi)}{p(D|\Theta)p(\Theta)}\right] + \log p(D)$$

Intractable log-marginal likelihood (does not depend on Psi)

The evidence lower bound (ELBO)

$$\mathrm{ELBO}(\Psi) = \mathbb{E}_{q(\Theta;\Psi)}\left[\log \frac{p(D|\Theta)p(\Theta)}{q(\Theta;\Psi)}\right]$$

Decomposition of the evidence lower bound

Averaged log-likelihood (data fit)

$$\mathrm{ELBO}(\Psi) = \mathbb{E}_{q(\Theta;\Psi)}[\log p(D|\Theta)] - D_{\mathrm{KL}}\left(q(\Theta|\Psi), p(\Theta)\right)$$

Tractable KL between approximation and prior (regularization)

A lower bound and approximation of the model evidence (much more of this in the VAE lecture!)

$$\mathrm{ELBO}(\Psi) \leq p(D)$$

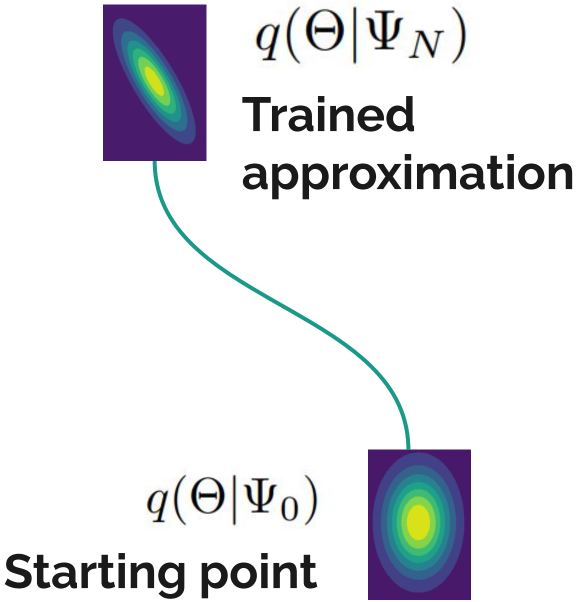
$$\mathrm{ELBO}(\Psi) = p(D) \text{ if } q(\Theta \mid \Psi) = p(\Theta \mid D)$$

Part III: Stochastic gradient estimation

Dr. Luca Ambrogioni

Bayesian inference by gradient descent of the negative ELBO

$$\Psi_{n+1} = \psi_n - \eta \nabla \left(-\mathrm{ELBO}(\Psi_n) \right)$$



Can we compute the gradient of the ELBO?

$$\nabla_\Psi \left(-\mathrm{ELBO}(\Psi_n)\right) = -\nabla_\Psi \mathbb{E}_{q(\Theta;\Psi)}\left[\log \frac{p(D|\Theta)p(\Theta)}{q(\Theta;\Psi)}\right]$$

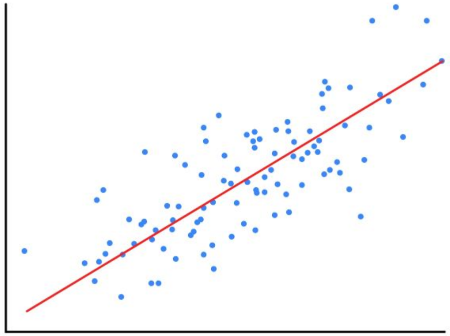
The variational parameters appear in two places

Can we move the gradient inside the expectation? (Nope!)

$$\nabla_\Psi \mathbb{E}_{q(\Theta;\Psi)}\left[\log \frac{p(D|\Theta)p(\Theta)}{q(\Theta;\Psi)}\right] \cancel{=} \mathbb{E}_{q(\Theta;\Psi)}\left[\nabla_\Psi \log \frac{p(D|\Theta)p(\Theta)}{q(\Theta;\Psi)}\right]$$

This misses the dependency on the expectation itself

Back to a very simple example!



Forward pass:

$$y_j = wx_j$$

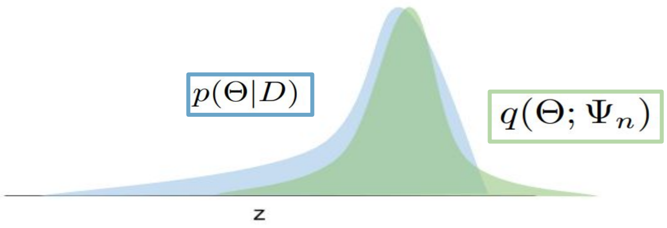
Likelihood:

$$\log p(D \mid w) = \sum_{j=1}^{N} \log p(y_j \mid x_j, w) = \sum_{j=1}^{N} \mathcal{L}_j(w)$$

Prior:

$$p(w) = \mathcal{N}(0, 1)$$

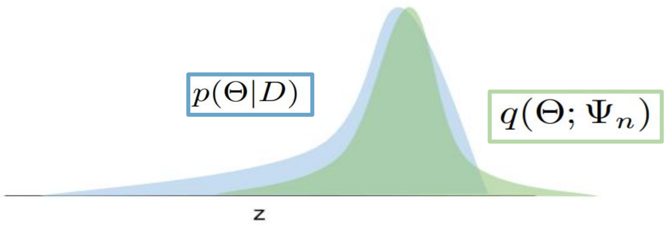
A simple variation model



$$q(w;\mu,\sigma) = \mathcal{N}\left(w;\mu,\sigma^2\right)$$

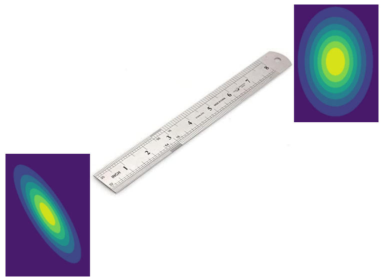
Variational parameters

Computing the ELBO



$$\mathrm{ELBO}(\mu,\sigma) = -\mathbb{E}_{q(w;\mu,\sigma)}[\log p(D|w)] + D_{\mathrm{KL}}\left(q(w|\mu,\sigma), \mathcal{N}(w;0,1)\right)$$

Computing the ELBO: The KL term



The KL divergence between two Gaussian distributions can be computed in closed form

$$D_{\mathrm{KL}}\left(q(w|\mu,\sigma),\mathcal{N}(w;0,1)\right) = -\log\sigma + \frac{1}{2}\left(\sigma^2+\mu^2\right) - \frac{1}{2}$$

Derivation:
https://stats.stackexchange.com/questions/7440/kl-divergence-between-two-univariate-gaussians

Computing the ELBO: Making progresses

We need to figure out how to take the gradient of this!

$$\mathrm{ELBO}(\mu, \sigma) = -\sum_{j=1}^{N} \mathbb{E}_{q(w;\mu,\sigma)}[\mathcal{L}_j(w)] - \log \sigma + \frac{1}{2}\left(\sigma^2 + \mu^2\right) - \frac{1}{2}$$

Easy to take the gradient of this!

Computing the gradients: The average likelihood term

$$\nabla_{\mu} \sum_{j=1}^{N} \mathbb{E}_{q(w;\mu,\sigma)}[\mathcal{L}_j(w)]$$

$$\nabla_{\sigma} \sum_{j=1}^{N} \mathbb{E}_{q(w;\mu,\sigma)}[\mathcal{L}_j(w)]$$

We would like to move the gradient inside the expectation. However, this would ignore the dependency on the parameters

Computing the gradients: The reparameterization trick

$$w = \sigma\epsilon + \mu$$

$$\epsilon \sim \mathcal{N}(w; 0, 1)$$

How to reparameterize: Express the variable as a deterministic transformation (dependent on the variational parameters) of a random variable that follows a fixed parameter-independent distribution.

Computing the gradients: The reparam trick

$$w = \sigma\epsilon + \mu$$

$$\epsilon \sim \mathcal{N}(w; 0, 1)$$

$$\nabla_\mu \sum_{j=1}^{N} \mathbb{E}_{q(w;\mu,\sigma)}[\mathcal{L}_j(w)] = \nabla_\mu \sum_{j=1}^{N} \mathbb{E}_{\mathcal{N}(\epsilon;0,1)}[\mathcal{L}_j(\sigma\epsilon + \mu)]$$

We can now move the gradient inside the expectation!

Computing the gradients: The reparam trick

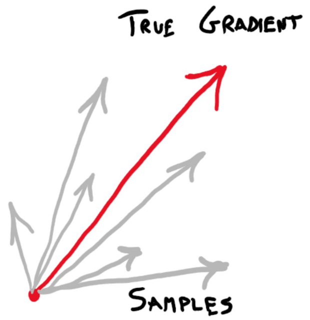
$$w = \sigma\epsilon + \mu$$

$$\epsilon \sim \mathcal{N}(w; 0, 1)$$

$$\nabla_\mu \sum_{j=1}^{N} \mathbb{E}_{\mathcal{N}(\epsilon;0,1)}[\mathcal{L}_j(\sigma\epsilon + \mu)] = \sum_{j=1}^{N} \mathbb{E}_{\mathcal{N}(\epsilon;0,1)}[\nabla_\mu \mathcal{L}_j(\sigma\epsilon + \mu)]$$

Just an average of gradients of regular randomized loss functions!

Stochastic gradient estimation



Unbiased gradient estimator

$$\sum_{j=1}^{N} \mathbb{E}_{\mathcal{N}(\epsilon;0,1)}[\nabla_\mu \mathcal{L}_j(\sigma\epsilon + \mu)] \approx \sum_{j=1}^{N} \frac{1}{M} \sum_{m=1}^{M} \nabla_\mu \mathcal{L}_j(\sigma\epsilon_m + \mu)$$

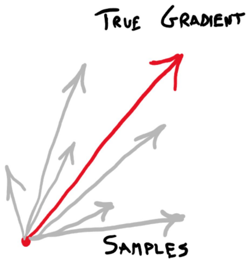
$$\epsilon_m \sim_{\text{iid}} \mathcal{N}(w; 0, 1)$$

Independently sampled random numbers
(A Monte Carlo method!)

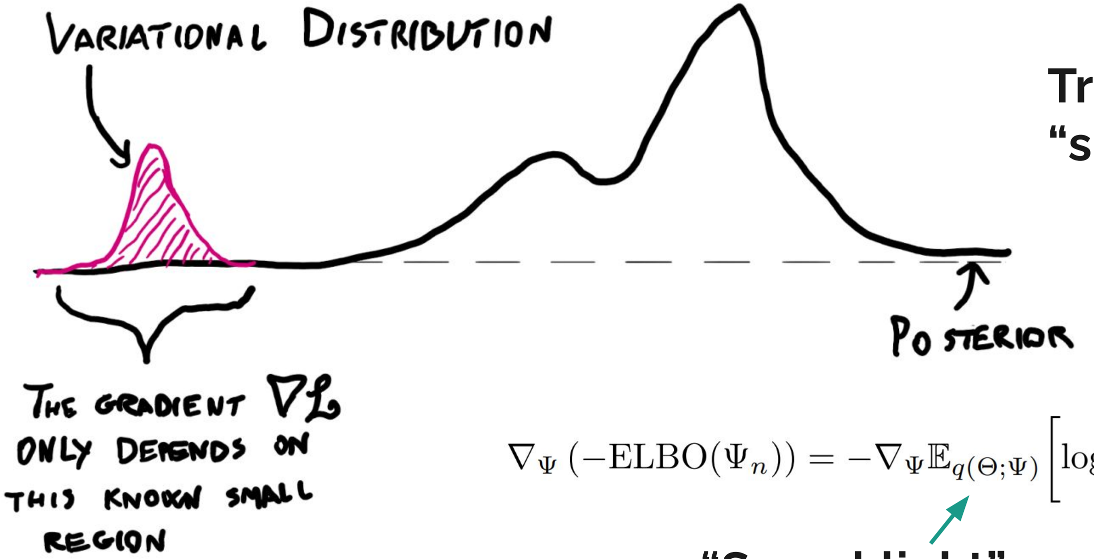
Computing the gradients: Putting everything together

$$\nabla_\mu \left(-\mathrm{ELBO}(\mu, \sigma)\right) \approx -\sum_{j=1}^{N} \frac{1}{M} \sum_{m=1}^{M} \nabla_\mu \mathcal{L}_j(\sigma \epsilon_m + \mu) - \nabla_\mu \left(-\log \sigma + \frac{1}{2}\left(\sigma^2 + \mu^2\right) \right)$$

$$= -\sum_{j=1}^{N} \frac{1}{M} \sum_{m=1}^{M} \nabla_\mu \mathcal{L}_j(\sigma \epsilon_m + \mu) - \mu$$



Variational gradient descent as a searchlight



True posterior
"shape"

$$\nabla_\Psi \left(-\mathrm{ELBO}(\Psi_n)\right) = -\nabla_\Psi \mathbb{E}_{q(\Theta;\Psi)}\left[\log \frac{p(D|\Theta)p(\Theta)}{q(\Theta;\Psi)}\right]$$

"Searchlight"

Thank you.