Practical Variational Inference for Neural Network (Alex Graves, NIPS 2011)

Basile Mayeur

GT-Deepnet reading group
October, 16th 2014

Classical setting for the neural network

- Choose w to fit the training database $\mathcal{D}$
  Do a gradient descent on the w to minimise the network loss:
  $L^N(w, \mathcal{D}) = -ln\ Pr(\mathcal{D}|w) = -\sum_{(x,y)\in\mathcal{D}} ln\ Pr(y|x, w)$
- Use w^* to predict $Pr(\mathcal{D}'|w^*)$

Classical setting for the neural network

- Choose w to fit the training database $\mathcal{D}$
  Do a gradient descent on the w to minimise the network loss:
  $L^N(w, \mathcal{D}) = -ln\ Pr(\mathcal{D}|w) = -\sum_{(x,y)\in\mathcal{D}} ln\ Pr(y|x, w)$
- Use w^* to predict $Pr(\mathcal{D}'|w^*)$

To avoid overfitting

Early stopping

Addition of a regularisation term to the loss function

- L1 regularisation: $\lambda\mathcal{N}_1(w)$
- L2 regularisation: $\lambda\mathcal{N}_2(w)$

Unsupervised pretraining

Dropout

Bayesian setting for a neural network

$$\underbrace{Pr(w|\mathcal{D},\alpha)}_{\text{Posterior distribution}} = \frac{\overbrace{Pr(\mathcal{D}|w)}^{\text{Likelihood}}\ \overbrace{P(w|\alpha)}^{\text{Prior}}}{\underbrace{\int_{w'} Pr(\mathcal{D}|w')P(w'|\alpha)}_{\text{Partition function}}}$$

Bayesian setting for a neural network

$$\underbrace{Pr(w|\mathcal{D},\alpha)}_{\text{Posterior distribution}} = \frac{\overbrace{Pr(\mathcal{D}|w)}^{\text{Likelihood}}\,\overbrace{P(w|\alpha)}^{\text{Prior}}}{\underbrace{\int_{w'} Pr(\mathcal{D}|w')P(w'|\alpha)}_{\text{Partition function}}}$$

Problem: The partition function is too complex to compute analytically or to sample from

Variational inference

Variational inference uses a simpler distribution $Q(w|\overbrace{\ \beta\ })$ to approximate $Pr(w|\mathcal{D}, \alpha)$

Variational inference

Variational inference uses a simpler distribution $Q(w|\overbrace{\;\beta\;})$ to approximate $Pr(w|\mathcal{D}, \alpha)$

$$\underset{w}{\text{Minimize:}} \quad \mathcal{F} = -\left\langle \ln\left(\frac{Pr(\mathcal{D}|w)P(w|\alpha)}{Q(w|\beta)}\right)\right\rangle_{w\sim Q(w|\beta)}$$

Variational inference

Variational inference uses a simpler distribution $Q(w|\overbrace{\ \beta\ })$ to approximate $Pr(w|\mathcal{D},\alpha)$

$$\underset{w}{\text{Minimize:}} \quad \mathcal{F} = -\Big\langle \ln\left(\frac{Pr(\mathcal{D}|w)P(w|\alpha)}{Q(w|\beta)}\right)\Big\rangle_{w\sim Q(w|\beta)}$$

$$\mathcal{F} \qquad = \quad \Big\langle -\ln(Pr(\mathcal{D}|w))\Big\rangle_{w\sim Q(w|\beta)} \quad + \quad \Big\langle -\ln\left(\tfrac{P(w|\alpha)}{Q(w|\beta)}\right)\Big\rangle_{w\sim Q(w|\beta)}$$

Variational inference

Variational inference uses a simpler distribution $Q(w|\overbrace{\beta})$ to approximate $Pr(w|\mathcal{D},\alpha)$

$$\underset{w}{\text{Minimize:}} \quad \mathcal{F} = -\left\langle \ln\left(\frac{Pr(\mathcal{D}|w)P(w|\alpha)}{Q(w|\beta)}\right)\right\rangle_{w\sim Q(w|\beta)}$$

$$\begin{aligned}
\mathcal{F} &= \left\langle -\ln(Pr(\mathcal{D}|w))\right\rangle_{w\sim Q(w|\beta)} + \left\langle -\ln\left(\frac{P(w|\alpha)}{Q(w|\beta)}\right)\right\rangle_{w\sim Q(w|\beta)} \\
\underbrace{\mathcal{F}}_{\text{Loss Function}} &= \underbrace{\left\langle L^N(w,\mathcal{D})\right\rangle_{w\sim Q(w|\beta)}}_{\text{Error Loss}} + \underbrace{D_{KL}(Q(.|\beta)||P(.|\alpha))}_{\text{Complexity Loss}}
\end{aligned}$$

Variational inference

Variational inference uses a simpler distribution $Q(w|\overbrace{\ \beta\ })$ to approximate $Pr(w|\mathcal{D},\alpha)$

$$\underset{w}{\text{Minimize:}} \quad \mathcal{F} = -\Big\langle \ln\left(\frac{Pr(\mathcal{D}|w)P(w|\alpha)}{Q(w|\beta)}\right)\Big\rangle_{w\sim Q(w|\beta)}$$

$$\begin{aligned}
\mathcal{F} &= \Big\langle -\ln(Pr(\mathcal{D}|w))\Big\rangle_{w\sim Q(w|\beta)} &+& \Big\langle -\ln\left(\tfrac{P(w|\alpha)}{Q(w|\beta)}\right)\Big\rangle_{w\sim Q(w|\beta)} \\
\underbrace{\mathcal{F}}_{\text{Loss Function}} &= \underbrace{\Big\langle L^N(w,\mathcal{D})\Big\rangle_{w\sim Q(w|\beta)}}_{\text{Error Loss}} &+& \underbrace{D_{KL}(Q(.|\beta)||P(.|\alpha))}_{\text{Complexity Loss}} \\
L(\alpha,\beta,\mathcal{D}) &= L^E(\beta,\mathcal{D}) &+& L^C(\alpha,\beta)
\end{aligned}$$

How to use $L(\alpha, \beta, \mathcal{D})$

1. Choose a class of distribution for $Q(.|\beta)$

How to use $L(\alpha, \beta, \mathcal{D})$

1. Choose a class of distribution for $Q(.|\beta)$
2. Choose the prior α

How to use $L(\alpha, \beta, \mathcal{D})$

1. Choose a class of distribution for $Q(.|\beta)$
2. Choose the prior α
3. Calculate the partial dervatives of the $L(\alpha, \beta, \mathcal{D})$ w.r.t β_i
   - Analytically for the simple cases
     or
   - Numerically using Monte Carlo sampling

How to use $L(\alpha, \beta, \mathcal{D})$

1. Choose a class of distribution for $Q(.|\beta)$
2. Choose the prior α
3. Calculate the partial dervatives of the $L(\alpha, \beta, \mathcal{D})$ w.r.t β_i
    - Analytically for the simple cases
      or
    - Numerically using Monte Carlo sampling
4. Update β_i

How to use $L(\alpha, \beta, \mathcal{D})$

1. Choose a class of distribution for $Q(.|\beta)$
2. Choose the prior α
3. Calculate the partial dervatives of the $L(\alpha, \beta, \mathcal{D})$ w.r.t β_i
   - Analytically for the simple cases
     or
   - Numerically using Monte Carlo sampling
4. Update β_i

In the next, we only considere diagonal posterior $Q(w|\beta) = \prod_{i=1}^{W} q_i(w_i|\beta_i)$ meaning that each w_i is sampled from $q_i(.|\beta_i)$.

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta) = \Big\langle -\ln\left(\frac{P(w|\alpha)}{Q(w|\beta)}\right) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta) = \Big\langle -\ln\left(\frac{P(w|\alpha)}{Q(w|\beta)}\right) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

If the prior is a uniform distribution:

$$\frac{\partial L^C(\alpha, \beta)}{\partial w_i} = 0$$

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta) = \Big\langle -\ln\left(\frac{P(w|\alpha)}{Q(w|\beta)}\right) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

If the prior is a uniform distribution:

$$\frac{\partial L^C(\alpha, \beta)}{\partial w_i} = 0$$

Uniform prior $\Rightarrow$ cost function: negative log-likelihood

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta) = \Big\langle -\ln\left(\frac{P(w|\alpha)}{Q(w|\beta)}\right) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle - \ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = - \ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta) = \Big\langle - \ln\Big(\frac{P(w|\alpha)}{Q(w|\beta)}\Big) \Big\rangle_{w \sim Q(w|\beta)} = - \ln P(w|\alpha) + C$$

if the prior is a Gaussian distribution, $\alpha = \{\mu, \sigma^2\}$:

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta) = \Big\langle -\ln\Big(\frac{P(w|\alpha)}{Q(w|\beta)}\Big) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

if the prior is a Gaussian distribution, $\alpha = \{\mu, \sigma^2\}$:

$$P(w|\alpha) = \prod_{i=1}^{W} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(w_i - \mu)^2}{2\sigma^2})$$

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta) = \Big\langle -\ln\Big(\frac{P(w|\alpha)}{Q(w|\beta)}\Big) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

if the prior is a Gaussian distribution, $\alpha = \{\mu, \sigma^2\}$:

$$P(w|\alpha) = \prod_{i=1}^{W} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(w_i - \mu)^2}{2\sigma^2})$$

$$\frac{\partial L^C(\alpha, \beta)}{\partial w_i} = \frac{w_i - \mu}{\sigma^2}$$

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta) = \Big\langle -\ln\Big(\frac{P(w|\alpha)}{Q(w|\beta)}\Big) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

if the prior is a Gaussian distribution, $\alpha = \{\mu, \sigma^2\}$:

$$P(w|\alpha) = \prod_{i=1}^{W} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(w_i - \mu)^2}{2\sigma^2})$$

$$\frac{\partial L^C(\alpha, \beta)}{\partial w_i} = \frac{w_i - \mu}{\sigma^2}$$

$\mu = 0$, σ^2 fixed $\Rightarrow$ L2 regularisation

If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w\sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta) = \Big\langle -\ln\Big(\frac{P(w|\alpha)}{Q(w|\beta)}\Big) \Big\rangle_{w\sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

if the prior is a Gaussian distribution, $\alpha = \{\mu, \sigma^2\}$:

$$P(w|\alpha) = \prod_{i=1}^{W} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(w_i - \mu)^2}{2\sigma^2})$$

$$\frac{\partial L^C(\alpha, \beta)}{\partial w_i} = \frac{w_i - \mu}{\sigma^2}$$

$\mu = 0$, σ^2 fixed $\Rightarrow$ L2 regularisation
Optimal prior $\hat{\alpha} = (\bar{w}, \frac{1}{W}\sum_{i=1}^{W}(w_i - \hat{\mu})^2)$

If $Q(.|\beta)$ is a gaussian distribution: $\beta_i = (\mu_i, \sigma_i^2)$

If $Q(.|\beta)$ is a gaussian distribution: $\beta_i = (\mu_i, \sigma_i^2)$

$$L^E(\beta, \mathcal{D}) = \left\langle L^N(w, \mathcal{D}) \right\rangle_{w \sim Q(w|\beta)} \approx -\frac{1}{S} \sum_{k=1}^{S} L^N(w^{(k)}, \mathcal{D})$$

If $Q(.|\beta)$ is a gaussian distribution: $\beta_i = (\mu_i, \sigma_i^2)$

$$L^E(\beta, \mathcal{D}) = \left\langle L^N(w, \mathcal{D}) \right\rangle_{w \sim Q(w|\beta)} \approx -\frac{1}{S} \sum_{k=1}^{S} L^N(w^{(k)}, \mathcal{D})$$

$$\frac{\partial L^E(\beta, \mathcal{D})}{\partial \mu_i} \approx -\frac{1}{S} \sum_{k=1}^{S} \frac{\partial L^N(w^{(k)}, \mathcal{D})}{\partial w_i}$$

If $Q(.|\beta)$ is a gaussian distribution: $\beta_i = (\mu_i, \sigma_i^2)$

$$L^E(\beta, \mathcal{D}) = \left\langle L^N(w, \mathcal{D}) \right\rangle_{w \sim Q(w|\beta)} \approx -\frac{1}{S} \sum_{k=1}^{S} L^N(w^{(k)}, \mathcal{D})$$

$$\frac{\partial L^E(\beta, \mathcal{D})}{\partial \mu_i} \approx -\frac{1}{S} \sum_{k=1}^{S} \frac{\partial L^N(w^{(k)}, \mathcal{D})}{\partial w_i}$$

$$\frac{\partial L^E(\beta, \mathcal{D})}{\partial \sigma_i^2} \approx -\frac{1}{2S} \sum_{k=1}^{S} \left[\frac{\partial L^N(w^{(k)}, \mathcal{D})}{\partial w_i} \right]^2$$

... and the prior is a gaussian distribution, $\alpha = \{\mu, \sigma^2\}$

$$L^C(\alpha, \beta) = \sum_{i=1}^{W} \ln \frac{\sigma}{\sigma_i} + \frac{1}{2\sigma^2} \left[(\mu_i - \mu)^2 + \sigma_i^2 - \sigma^2 \right]$$

... and the prior is a gaussian distribution, $\alpha = \{\mu, \sigma^2\}$

$$L^C(\alpha, \beta) = \sum_{i=1}^{W} \ln \frac{\sigma}{\sigma_i} + \frac{1}{2\sigma^2}\left[(\mu_i - \mu)^2 + \sigma_i^2 - \sigma^2\right]$$

$$\frac{\partial L^C(\alpha, \beta)}{\partial \mu_i} = \frac{\mu_i - \mu}{\sigma^2} \quad , \quad \frac{\partial L^C(\alpha, \beta))}{\partial \sigma_i^2} = \frac{1}{2}\left[\frac{1}{\sigma^2} - \frac{1}{\sigma_i^2}\right]$$

... and the prior is a gaussian distribution, $\alpha = \{\mu, \sigma^2\}$

$$L^C(\alpha, \beta) = \sum_{i=1}^{W} \ln \frac{\sigma}{\sigma_i} + \frac{1}{2\sigma^2} \left[(\mu_i - \mu)^2 + \sigma_i^2 - \sigma^2 \right]$$

$$\frac{\partial L^C(\alpha, \beta)}{\partial \mu_i} = \frac{\mu_i - \mu}{\sigma^2} \quad , \quad \frac{\partial L^C(\alpha, \beta))}{\partial \sigma_i^2} = \frac{1}{2} \left[\frac{1}{\sigma^2} - \frac{1}{\sigma_i^2} \right]$$

Optimal prior $\hat{\alpha} = \left(\frac{1}{W} \sum_{i=1}^{W} \mu_i \ , \ \frac{1}{W} \sum_{i=1}^{W} \left[\sigma_i^2 + (\mu_i - \hat{\mu})^2 \right] \right)$

In practice

In practice

For the learning

Choose the type of distribution $Q(., \beta)$
Initialise β
Choose the type of the priors
loop
  Compute the optimal prior $\hat{\alpha}$ from β
  $w \sim Q(.|\beta)$
  Take (x, y) from the training base
  $\forall k$, compute $\frac{\partial L(\hat{\alpha},\beta,(x,y))}{\partial \beta_k}$
  Update β
end loop

In practice

For the learning

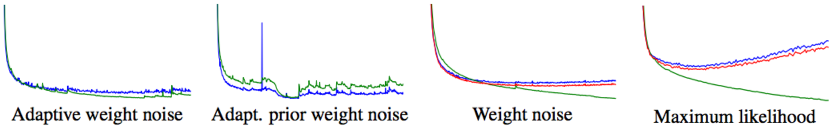
Choose the type of distribution $Q(., \beta)$
Initialise β
Choose the type of the priors
loop
  Compute the optimal prior $\hat{\alpha}$ from β
  $w \sim Q(.|\beta)$
  Take (x, y) from the training base
  $\forall k$, compute $\frac{\partial L(\hat{\alpha},\beta,(x,y))}{\partial \beta_k}$
  Update β
end loop

For the prediction

$\forall i, w_i =< w_i >_{w_i \sim q(.|\beta_{i*})}$

Experimentation on the TIMIT database

Name	Posterior	Prior	Error	Epochs
Adaptive L1	Delta	Laplace	49.0	7
Adaptive L2	Delta	Gauss	35.1	421
Adaptive mean L2	Delta	Gauss $\sigma^2 = 0.1$	28.0	53
L2	Delta	Gauss $\mu = 0, \sigma^2 = 0.1$	27.4	59
Maximum likelihood	Delta	Uniform	27.1	44
L1	Delta	Laplace $\mu = 0, b = 1/12$	26.0	545
Adaptive mean L1	Delta	Laplace $b = 1/12$	25.4	765
Weight noise	Gauss $\sigma_i = 0.075$	Uniform	25.4	220
Adaptive prior weight noise	Gauss $\sigma_i = 0.075$	Gauss	24.7	260
Adaptive weight noise	Gauss	Gauss	23.8	384



Main contributions of the article

- Apply the Bayesian model associated to the variational inference to generalize the common loss functions of the neural network
- Find a practical way to apply variational inference on the posterior using the Monte-Carlo sampling
- The neural network doesn't overfit with the gaussian posterior loss function
- Obtain state of the art performance for a shallow recurrent neural net on the TIMIT database

Annexe: If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta)) = \Big\langle -\ln\left(\frac{P(w|\alpha)}{Q(w|\beta)}\right) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

Annexe: If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w))\Big\rangle_{w\sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta)) = \Big\langle -\ln\left(\frac{P(w|\alpha)}{Q(w|\beta)}\right)\Big\rangle_{w\sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

if the prior is a Laplace distribution, $\alpha = \{\mu, b\}$:

Annexe: If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta)) = \Big\langle -\ln\Big(\frac{P(w|\alpha)}{Q(w|\beta)}\Big) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

if the prior is a Laplace distribution, $\alpha = \{\mu, b\}$:

$$P(w|\alpha) = \prod_{i=1}^{W} \frac{1}{2b} \exp(-\frac{|w_i - \mu|}{b})$$

Annexe: If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta)) = \Big\langle -\ln\left(\frac{P(w|\alpha)}{Q(w|\beta)}\right) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

if the prior is a Laplace distribution, $\alpha = \{\mu, b\}$:

$$P(w|\alpha) = \prod_{i=1}^{W} \frac{1}{2b} \exp(-\frac{|w_i - \mu|}{b})$$

$$\frac{\partial L^C(\alpha, \beta))}{\partial w_i} = \frac{sgn(w_i - \mu)}{b}$$

Annexe: If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w \sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta)) = \Big\langle -\ln\Big(\frac{P(w|\alpha)}{Q(w|\beta)}\Big) \Big\rangle_{w \sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

if the prior is a Laplace distribution, $\alpha = \{\mu, b\}$:

$$P(w|\alpha) = \prod_{i=1}^{W} \frac{1}{2b} \exp(-\frac{|w_i - \mu|}{b})$$

$$\frac{\partial L^C(\alpha, \beta))}{\partial w_i} = \frac{sgn(w_i - \mu)}{b}$$

$\mu = 0$, b fixed $\Rightarrow$ L1 regularisation

Annexe: If $Q(.|\beta)$ is a delta distribution: $\beta_i = w_i$

$$L^E(\beta, \mathcal{D}) = \Big\langle -\ln(Pr(\mathcal{D}|w)) \Big\rangle_{w\sim Q(w|\beta)} = -\ln(Pr(\mathcal{D}|w))$$

$$L^C(\alpha, \beta)) = \Big\langle -\ln\Big(\frac{P(w|\alpha)}{Q(w|\beta)}\Big) \Big\rangle_{w\sim Q(w|\beta)} = -\ln P(w|\alpha) + C$$

if the prior is a Laplace distribution, $\alpha = \{\mu, b\}$:

$$P(w|\alpha) = \prod_{i=1}^{W} \frac{1}{2b} \exp(-\frac{|w_i - \mu|}{b})$$

$$\frac{\partial L^C(\alpha, \beta))}{\partial w_i} = \frac{sgn(w_i - \mu)}{b}$$

$\mu = 0$, b fixed $\Rightarrow$ L1 regularisation
Optimal prior $\hat{\alpha} = (\mathrm{median(w)}, \frac{1}{W}\sum_{i=1}^{W} |w_i - \hat{\mu}|)$