Dropout as a Bayesian Approximation

Ivan Rodriguez

Main reference papers:

1 - “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”, Y.Gal and Z.Ghahramani – ICML’16 ~ 6400 citations

2 - “Dropout as a Bayesian Approximation: Appendix”, Y.Gal and Z.Ghahramani - ICML’16

Seminar outline:

1 - Standard Dropout Introduction

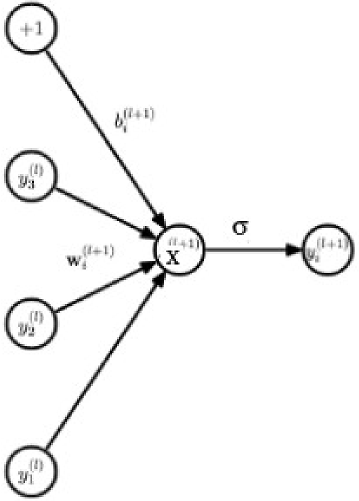
2 - Gaussian Process for DNN

3 - Dropout from a Bayesian point of view

4 - Results

1 - Standard Dropout Introduction

Original dropout method

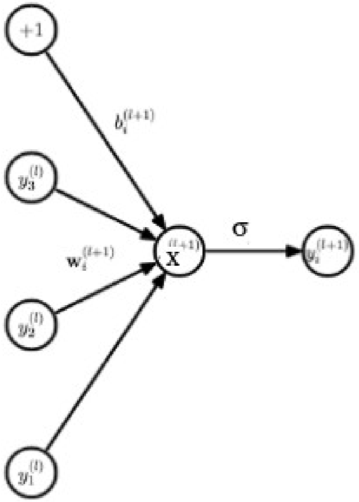


(a) Standard network

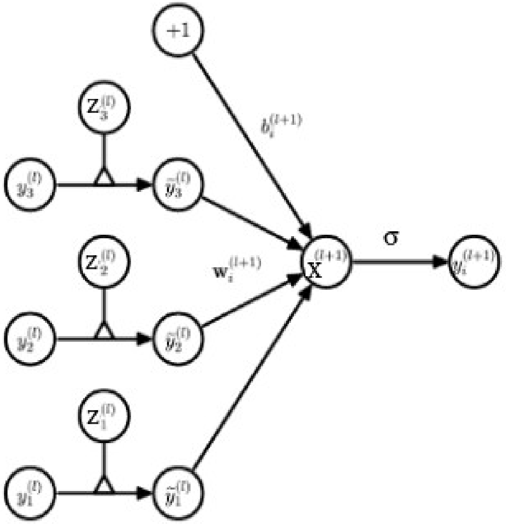
$$x^{(l+1)} = W^{(l+1)} y^{(l)} + b^{(l+1)}$$

$$y^{(l+1)} = \sigma(x^{(l+1)})$$

Original dropout method



(a) Standard network



(b) Dropout network

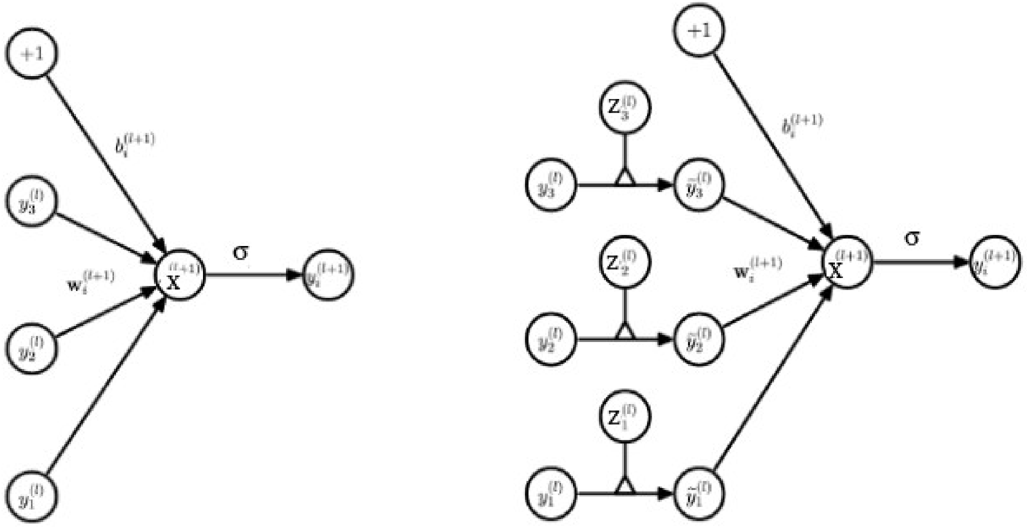
$$z^{(l)} = (z_1, z_2, \ldots) \text{ with } z_i \sim Bernoulli(p)$$

$$\tilde{y}^{(l)} = z^{(l)} \odot y^{(l)} = (z_1^{(l)} y_1^{(l)}, z_2^{(l)} y_2^{(l)}, \ldots)$$

$$x^{(l+1)} = W^{(l+1)} y^{(l)} + b^{(l+1)}$$

$$y^{(l+1)} = \sigma(x^{(l+1)})$$

Original dropout method



(a) Standard network (b) Dropout network

$$z^{(l)} = (z_1, z_2, \ldots) \text{ with } z_i \sim Bernoulli(p)$$

$$\tilde{y}^{(l)} = z^{(l)} \odot y^{(l)} = (z_1^{(l)} y_1^{(l)}, z_2^{(l)} y_2^{(l)}, \ldots)$$

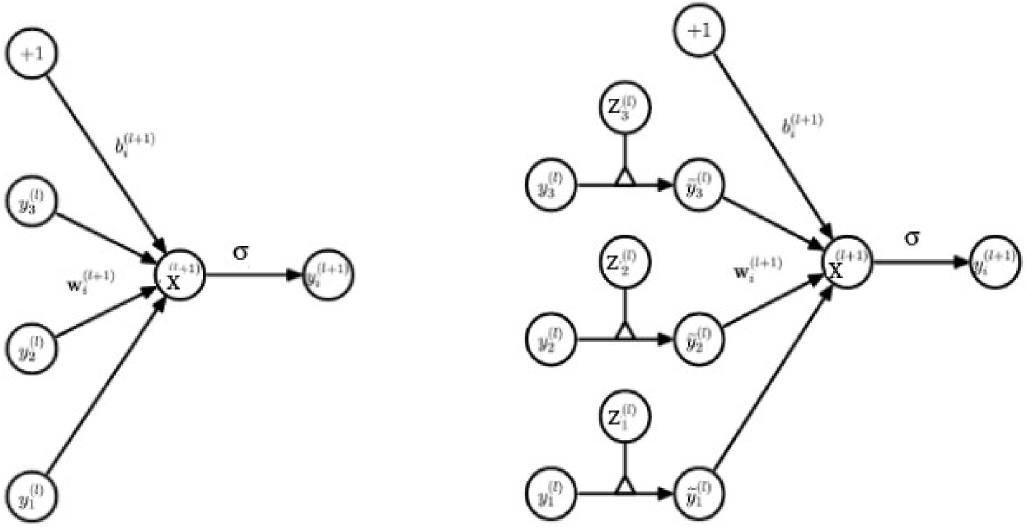
$$x^{(l+1)} = W^{(l+1)} y^{(l)} + b^{(l+1)}$$

$$y^{(l+1)} = \sigma(x^{(l+1)})$$

$$x^{(l+1)} = W^{(l+1)} \tilde{y}^{(l)} + b^{(l+1)} = \boxed{W^{(l+1)} diag(z^{(l)})} \; y^{(l)} + b^{(l+1)}$$

$$y^{(l+1)} = \sigma(x^{(l+1)})$$

Original dropout method



(a) Standard network (b) Dropout network

$$x^{(l+1)} = W^{(l+1)} y^{(l)} + b^{(l+1)}$$

$$y^{(l+1)} = \sigma(x^{(l+1)})$$

$z^{(l)} = (z_1, z_2, \ldots)$ with $z_i \sim Bernoulli(p)$

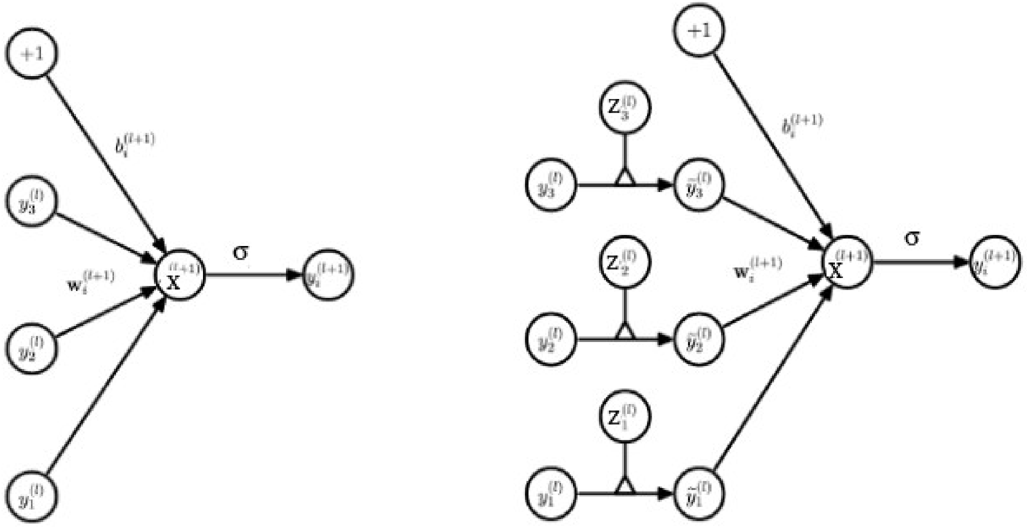
$$\tilde{y}^{(l)} = z^{(l)} \odot y^{(l)} = (z_1^{(l)} y_1^{(l)}, z_2^{(l)} y_2^{(l)}, ..)$$

$$x^{(l+1)} = W^{(l+1)} \tilde{y}^{(l)} + b^{(l+1)} = \boxed{W^{(l+1)} diag(z^{(l)})}\; y^{(l)} + b^{(l+1)}$$

$$y^{(l+1)} = \sigma(x^{(l+1)})$$

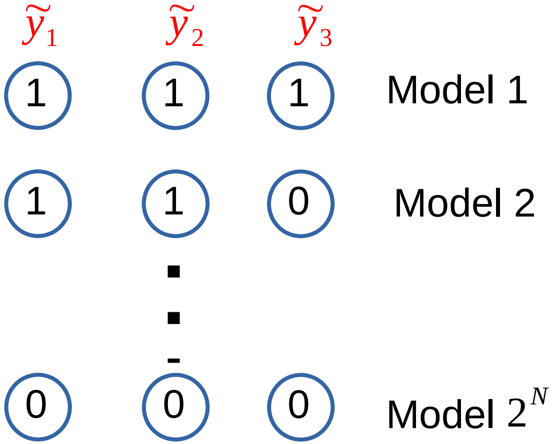
$\tilde{W}_{\alpha\beta}^{(l+1)} = W_{\alpha\beta}^{(l+1)} z_\beta^{(l)} = 0$ if $z_\beta = 0$ column β vanishes

Original dropout method



(a) Standard network (b) Dropout network

Dropout interpretation: Ensemble model



$$z^{(l)}=(z_1,z_2,\ldots)\ \text{with}\ z_i \sim Bernoulli(p)$$

$$\tilde{y}^{(l)}=z^{(l)}\odot y^{(l)}=(z_1^{(l)}y_1^{(l)},z_2^{(l)}y_2^{(l)},..)$$

$$x^{(l+1)}=W^{(l+1)}y^{(l)}+b^{(l+1)}$$

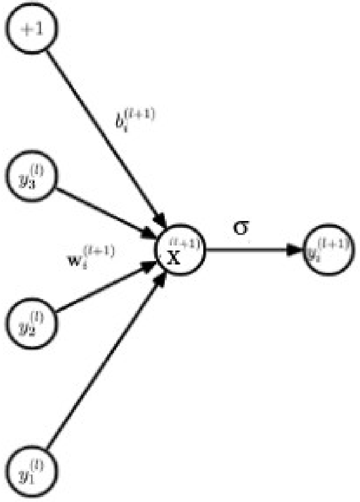
$$y^{(l+1)}=\sigma(x^{(l+1)})$$

$$x^{(l+1)}=W^{(l+1)}\tilde{y}^{(l)}+b^{(l+1)}=\boxed{W^{(l+1)}diag(z^{(l)})}\ y^{(l)}+b^{(l+1)}$$

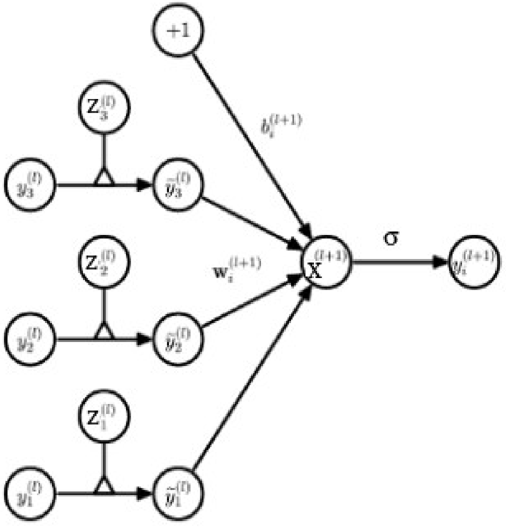
$$y^{(l+1)}=\sigma(x^{(l+1)})$$

$$\tilde{W}_{\alpha\beta}^{(l+1)}=W_{\alpha\beta}^{(l+1)}z_\beta^{(l)}=0\ \text{if}\ z_\beta=0$$ column β vanishes

Original dropout method



(a) Standard network



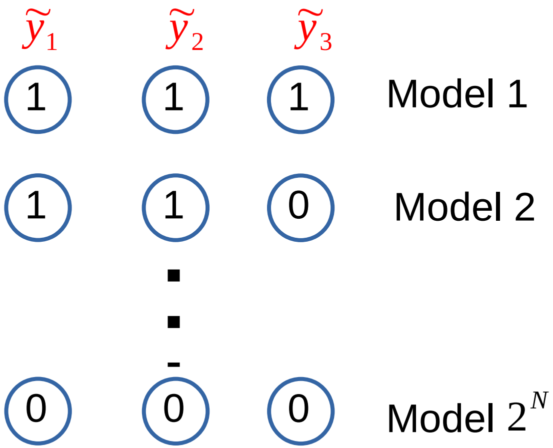
(b) Dropout network

$$x^{(l+1)} = W^{(l+1)} y^{(l)} + b^{(l+1)}$$
$$y^{(l+1)} = \sigma(x^{(l+1)})$$

$$z^{(l)} = (z_1, z_2, \dots) \text{ with } z_i \sim Bernoulli(p)$$
$$\tilde{y}^{(l)} = z^{(l)} \odot y^{(l)} = (z_1^{(l)} y_1^{(l)}, z_2^{(l)} y_2^{(l)}, \dots)$$
$$x^{(l+1)} = W^{(l+1)} \tilde{y}^{(l)} + b^{(l+1)} = \boxed{W^{(l+1)} diag(z^{(l)})} \; y^{(l)} + b^{(l+1)}$$
$$y^{(l+1)} = \sigma(x^{(l+1)})$$

$$\tilde{W}_{\alpha\beta}^{(l+1)} = W_{\alpha\beta}^{(l+1)} z_\beta^{(l)} = 0 \text{ if } z_\beta = 0$$ column β vanishes

Dropout interpretation: Ensemble model

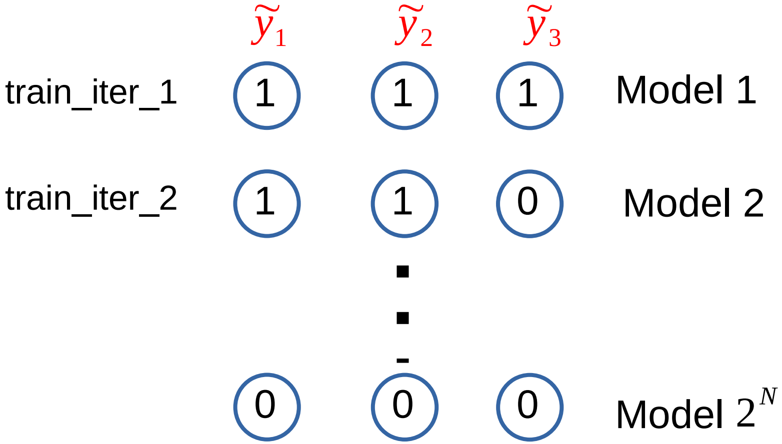


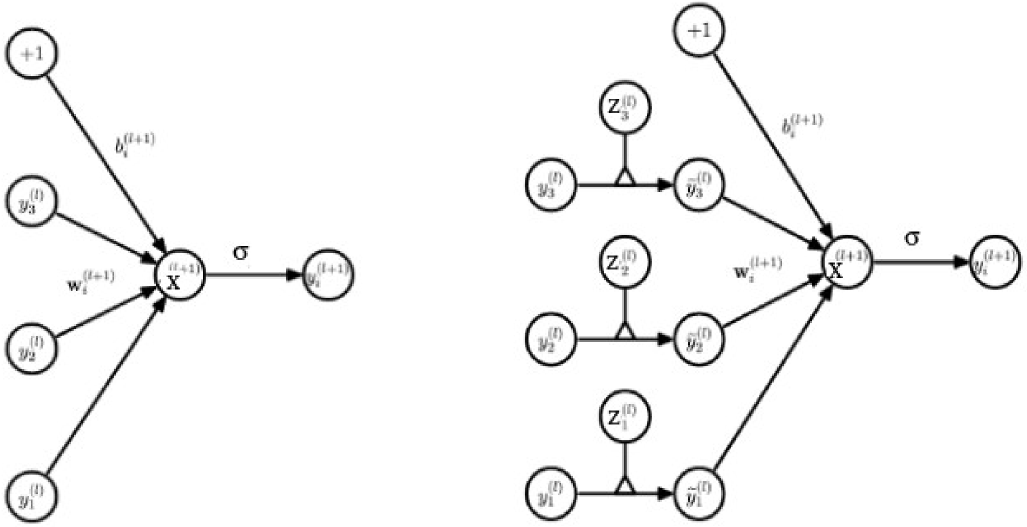
Inference:

$$E[\tilde{y}_i] = E[z_i y_i] = p\, y_i + (1-p)0 = p\, y_i$$
$$E[x_i] = W_{ij}\, p\, y_j^{(l)} + b_i \quad \Rightarrow \quad \boxed{W \to p\,W}$$

Original dropout method

Dropout interpretation: Ensemble model





(a) Standard network

(b) Dropout network

Obviously this interpretation is an approximation as the models are not really independent !

$$z^{(l)} = (z_1, z_2, \ldots) \text{ with } z_i \sim Bernoulli(p)$$

$$\tilde{y}^{(l)} = z^{(l)} \odot y^{(l)} = (z_1^{(l)} y_1^{(l)}, z_2^{(l)} y_2^{(l)}, \ldots)$$

$$x^{(l+1)} = W^{(l+1)} \tilde{y}^{(l)} + b^{(l+1)} = \boxed{W^{(l+1)} diag(z^{(l)})}\; y^{(l)} + b^{(l+1)}$$

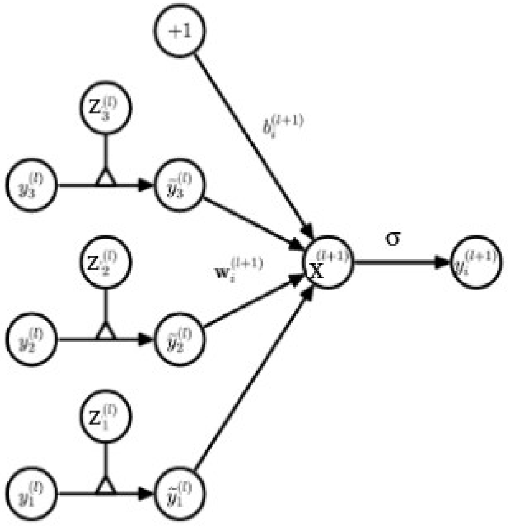
$$y^{(l+1)} = \sigma(x^{(l+1)})$$

$$\tilde{W}_{\alpha\beta}^{(l+1)} = W_{\alpha\beta}^{(l+1)} z_\beta^{(l)} = 0 \text{ if } z_\beta = 0 \quad \text{column } \beta \text{ vanishes}$$

$$x^{(l+1)} = W^{(l+1)} y^{(l)} + b^{(l+1)}$$

$$y^{(l+1)} = \sigma(x^{(l+1)})$$

Original dropout method



(b) Dropout network

In summary (for a 2-layer DNN): $\text{Data}: D=(X,Y)=\{(x_i,y_i) \quad i=1,\dots,N\}$

$x_i \sim 1\times Q;\ y_i \sim 1\times D$

$$\hat{y}=W^{(2)}diag(z^{(2)})\,\sigma(W^{(1)}diag(z^{(1)})x+b)$$

$W^{(1)}\sim Q\times K;\ W^{(2)}\sim D\times K$

$z^{(l)}=(z_1^{(l)},z_2^{(l)},\dots,z_Q^{(l)})$ with $z_j^{(l)}\sim Bernoulli(p)$

$$E=\frac{1}{2N}\sum_{n=1}^{N}\|y_n-\hat{y}_n\|$$

$$L_{dropout}=E+\lambda_1\|W_1\|+\lambda_2\|W_2\|+\lambda_3\|b\|$$

Training:

MLE with sampling $z^{(l)}\sim Bern.(p)$

Inference:

$W\rightarrow pW$

Using the Bayesian approach the ensemble picture would be more clear and we will recover the equations above and more !! .

2 - Gaussian Process for DNN

Gaussian process: short introduction

Given some observed data: $D = \{X, Y\}$

$$P(\hat{y}/D, \hat{x}) = \int df \ \ P(\hat{y}/f, \hat{x}, X) \boxed{P(f/D)}$$

→ Very hard to compute in gral. !!

Gaussian process: short introduction

Given some observed data: $D = \{X, Y\}$

$$P(\hat{y}/D, \hat{x}) = \int df \;\; P(\hat{y}/f, \hat{x}, X) \boxed{P(f/D)}$$ → Very hard to compute in gral. !!

In a Bayesian approach we use the **Bayes theorem** to get the posterior :

$$P(f/D) = \frac{P(Y/f, X)\, P(f/X)}{P(Y/X)}$$

Likelihood

$P(Y/f, X)$ → $N(y, f, \sigma^2)$

Prior

$P(f/X) = N(f, 0, K(X, X'))$

GP with kernel K

Evidence or Normalization

$$P(Y/X) = \int df\, P(Y, f/X) = \int df\, P(Y/f, X)\, P(f/X)$$

Gaussian process: short introduction

Given some observed data: $D=\{X,Y\}$

$$P(\hat{y}/D,\hat{x})=\int df \; P(\hat{y}/f,\hat{x},X)\boxed{P(f/D)}$$ → Very hard to compute in gral. !!

In a Bayesian approach we use the **Bayes theorem** to get the posterior :

$$P(f/D)=\frac{P(Y/f,X)P(f/X)}{P(Y/X)}$$

Likelihood

$P(Y/f,X)$ → $N(y,f,\sigma^2)$

Prior

$P(f/X)=N(f,0,K(X,X'))$

GP with kernel K

Evidence or Normalization

$$P(Y/X)=\int df\, P(Y,f/X)=\int df\, P(Y/f,X)P(f/X)$$

Kernels: include info about family of functions being approximated, i.e. periodic functions, smooth functions, stochastic functions, etc. as well as information about the confidence intervals. See link .

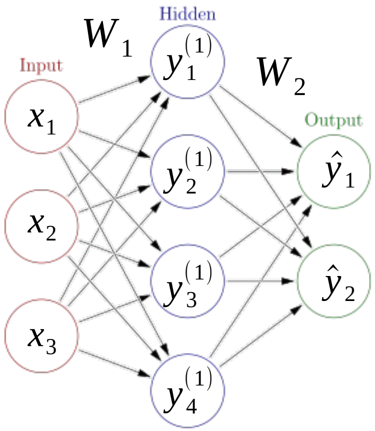
Gaussian process: Bayesian DNN connection

I will focus in a two layer DNN for simplicity. Suppose we have this kernel:

$$K(x,x')=\int p(w)\,p(b)\,\sigma(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}+b)^T\sigma(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}'+b)\,dw\,db$$

$w,x \sim Q\times 1 \qquad b\in\mathbb{R}$

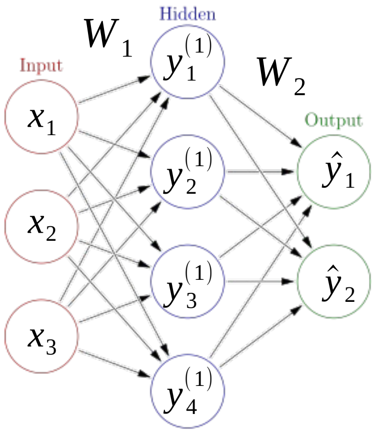
σ: ReLU, sigmoid, etc.



Gaussian process: Bayesian DNN connection

I will focus in a two layer DNN for simplicity. Suppose we have this kernel:

σ: ReLU, sigmoid, etc.



$$K(x, x') = \int p(w)\, p(b)\, \sigma(\mathbf{w}^T \mathbf{x} + b)^T \sigma(\mathbf{w}^T \mathbf{x}' + b)\, dw\, db$$

$w, x \sim Q \times 1 \qquad b \in \mathbb{R}$

Approximation: $\int \rightarrow \sum$

$$K(x, x') \simeq \sum_{k=1}^{K} \sqrt{\frac{1}{K}}\, \sigma(\mathbf{w}_k^T \mathbf{x} + b_k)^T \sqrt{\frac{1}{K}}\, \sigma(\mathbf{w}_k^T \mathbf{x}' + b_k)$$

$w_k, b_k \sim p(w), p(b)$

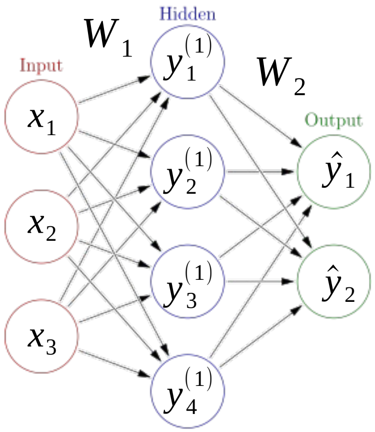
Gaussian process: Bayesian DNN connection

I will focus in a two layer DNN for simplicity. Suppose we have this kernel:

$$K(x,x')=\int p(w)\,p(b)\,\sigma(\boldsymbol{w}^T\boldsymbol{x}+b)^T\,\sigma(\boldsymbol{w}^T\boldsymbol{x}'+b)\,dw\,db$$

$w,x\sim Q\times 1 \qquad b\in\mathbb{R}$

σ: ReLU, sigmoid, etc.



Approximation: $\int \rightarrow \sum$

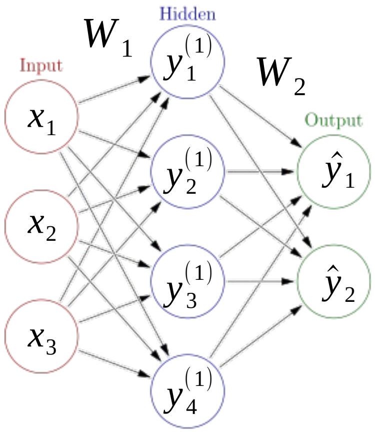
$$K(x,x')\simeq\sum_{k=1}^{K}\sqrt{\frac{1}{K}}\,\sigma(\boldsymbol{w}_k^T\boldsymbol{x}+b_k)^T\sqrt{\frac{1}{K}}\,\sigma(\boldsymbol{w}_k^T\boldsymbol{x}'+b_k)=\sqrt{\frac{1}{K}}\,\sigma(\boldsymbol{W}_1\boldsymbol{x}+b)^T\sqrt{\frac{1}{K}}\,\sigma(\boldsymbol{W}_1\boldsymbol{x}'+b)$$

$w_k,b_k\sim p(w),p(b) \qquad W_1\sim K\times Q$

Gaussian process: Bayesian DNN connection

I will focus in a two layer DNN for simplicity. Suppose we have this kernel:

σ: ReLU, sigmoid, etc.



$$K(x,x') = \int p(w)\,p(b)\,\sigma(\mathbf{w}^T\mathbf{x}+b)^T\,\sigma(\mathbf{w}^T\mathbf{x}'+b)\,dw\,db$$

$w, x \sim Q \times 1 \qquad b \in \mathbb{R}$

Approximation: $\int \rightarrow \sum$

$$K(x,x') \simeq \sum_{k=1}^{K} \sqrt{\frac{1}{K}}\,\sigma(\mathbf{w}_k^T\mathbf{x}+b_k)^T \sqrt{\frac{1}{K}}\,\sigma(\mathbf{w}_k^T\mathbf{x}'+b_k) = \sqrt{\frac{1}{K}}\,\underbrace{\sigma(\mathbf{W}_1\mathbf{x}+b)^T}_{\mathbf{y}^{(1)}} \sqrt{\frac{1}{K}}\,\sigma(\mathbf{W}_1\mathbf{x}'+b)$$

$w_k, b_k \sim p(w), p(b) \qquad W_1 \sim K \times Q$

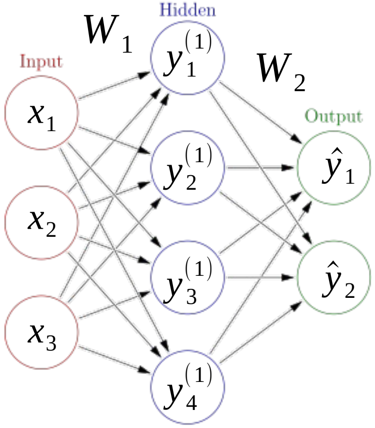
Gaussian process: Bayesian DNN connection

$$K(x,x')=\sqrt{\frac{1}{K}}\sigma(W_1x+b)^T\sqrt{\frac{1}{K}}\sigma(W_1x'+b)\sim y^{(1)}(W_1,x,b)^T y^{(1)}(W_1,x',b)$$

$$W_1\sim K\times Q;\ \ x\sim Q\times 1;\ \ y^{(1)}\sim K\times 1$$

with W_1 and b random variables as usual in a Bayesian approach

σ: ReLU, sigmoid, etc.



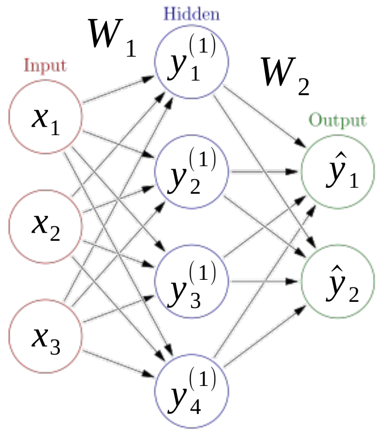
Gaussian process: Bayesian DNN connection

$$K(x,x')=\sqrt{\frac{1}{K}}\sigma(W_1 x+b)^T\sqrt{\frac{1}{K}}\sigma(W_1 x'+b)\sim y^{(1)}(W_1,x,b)^T y^{(1)}(W_1,x',b)$$

$$W_1\sim K\times Q;\;\; x\sim Q\times 1;\;\; y^{(1)}\sim K\times 1$$

with W_1 and b random variables as usual in a Bayesian approach

σ: ReLU, sigmoid, etc.



To find the connection between a GP and DNN's let's make use of the evidence:

$$P(Y/X)=\int P(Y/f)P(f/W_1,b,X)P(W_1)P(b)\,df\,dW_1\,db$$

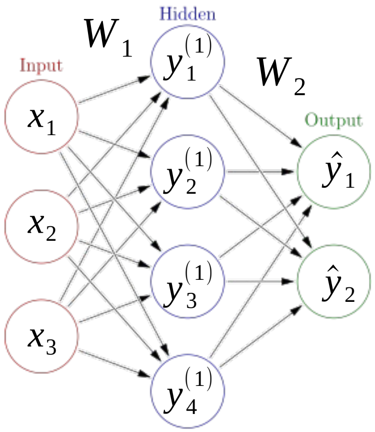
Gaussian process: Bayesian DNN connection

$$K(x,x')=\sqrt{\frac{1}{K}}\sigma(W_1x+b)^T\sqrt{\frac{1}{K}}\sigma(W_1x'+b)\sim y^{(1)}(W_1,x,b)^T y^{(1)}(W_1,x',b)$$

$$W_1\sim K\times Q;\ \ x\sim Q\times 1;\ \ y^{(1)}\sim K\times 1$$

with W_1 and b random variables as usual in a Bayesian approach

σ: ReLU, sigmoid, etc.



To find the connection between a GP and DNN's let's make use of the evidence:

$$P(Y/X)=\int P(Y/f)P(f/W_1,b,X)P(W_1)P(b)df\,dW_1db$$

$P(Y/f)\sim N(Y,f,\tau^{-1}I_D)$ Likelihood

$P(f/W_1,b,X)\sim GP(0,K)$

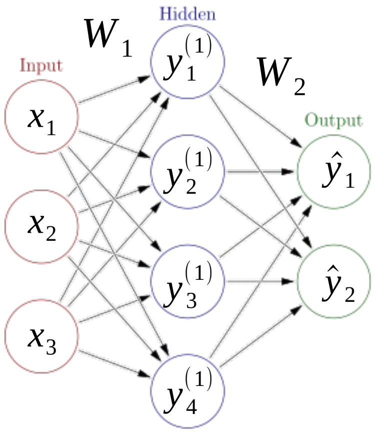
Gaussian process: Bayesian DNN connection

$$K(x,x')=\sqrt{\frac{1}{K}}\sigma(W_1x+b)^T\sqrt{\frac{1}{K}}\sigma(W_1x'+b)\sim y^{(1)}(W_1,x,b)^T y^{(1)}(W_1,x',b)$$

$$W_1\sim K\times Q;\ \ x\sim Q\times 1;\ \ y^{(1)}\sim K\times 1$$

with W_1 and b random variables as usual in a Bayesian approach

σ: ReLU, sigmoid, etc.



To find the connection between a GP and DNN's let's make use of the evidence:

$$P(Y/X)=\int P(Y/f)P(f/W_1,b,X)P(W_1)P(b)df\,dW_1db$$

$P(Y/f)\sim N(Y,f,\tau^{-1}I_D)$ Likelihood

$P(f/W_1,b,X)\sim GP(0,K)$

$$\int df\,P(Y/f)P(f/W_1,b,X)=N(Y;0,(K(X,X)+\tau^{-1})I_D)$$

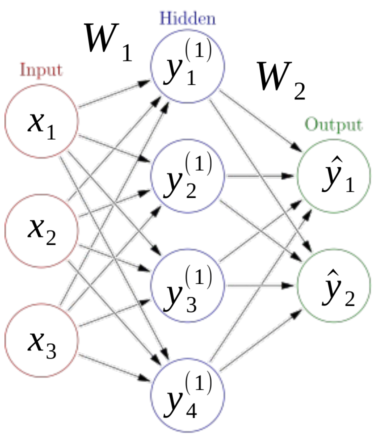
Gaussian process: Bayesian DNN connection

$$K(x,x')=\sqrt{\frac{1}{K}}\sigma(W_1x+b)^T\sqrt{\frac{1}{K}}\sigma(W_1x'+b)\sim y^{(1)}(W_1,x,b)^T y^{(1)}(W_1,x',b)$$

$$W_1\sim K\times Q;\;\; x\sim Q\times 1;\;\; y^{(1)}\sim K\times 1$$

with W_1 and b random variables as usual in a Bayesian approach

σ: ReLU, sigmoid, etc.



To find the connection between a GP and DNN's let's make use of the evidence:

$$P(Y/X)=\int P(Y/f)P(f/W_1,b,X)P(W_1)P(b)df\,dW_1 db$$

$P(Y/f)\sim N(Y,f,\tau^{-1}I_D)$ Likelihood

$P(f/W_1,b,X)\sim GP(0,K)$

$$\int df\,P(Y/f)P(f/W_1,b,X)=N(Y;0,(K(X,X)+\tau^{-1})I_D)=\int N(Y;W_2y^{(1)},\tau^{-1}I_D)P(W_2)dW_2$$

auxiliar matrix variables $:W_2$

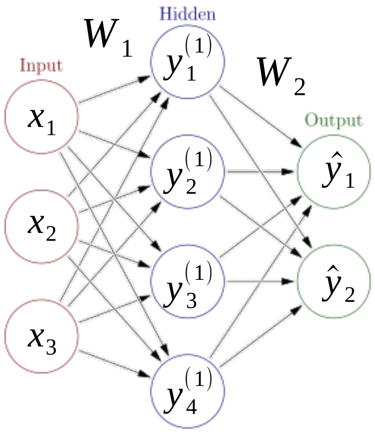
Gaussian process: Bayesian DNN connection

$$K(x,x')=\sqrt{\frac{1}{K}}\sigma(W_1x+b)^T\sqrt{\frac{1}{K}}\sigma(W_1x'+b)\sim y^{(1)}(W_1,x,b)^T y^{(1)}(W_1,x',b)$$

$$W_1\sim K\times Q;\; x\sim Q\times 1;\; y^{(1)}\sim K\times 1$$

with W_1 and b random variables as usual in a Bayesian approach

σ: ReLU, sigmoid, etc.



To find the connection between a GP and DNN's let's make use of the evidence:

$$P(Y/X)=\int P(Y/f)P(f/W_1,b,X)P(W_1)P(b)\,df\,dW_1\,db$$

$P(Y/f)\sim N(Y,f,\tau^{-1}I_D)$ Likelihood

$P(f/W_1,b,X)\sim GP(0,K)$

$$\int df\,P(Y/f)P(f/W_1,b,X)=N(Y;0,(K(X,X)+\tau^{-1})I_D)=\int N(Y;W_2y^{(1)},\tau^{-1}I_D)P(W_2)dW_2$$

auxiliar matrix variables : W_2

$$f(X,W_1,W_2,b)=\hat{y}=W_2\sigma(W_1X+b)$$

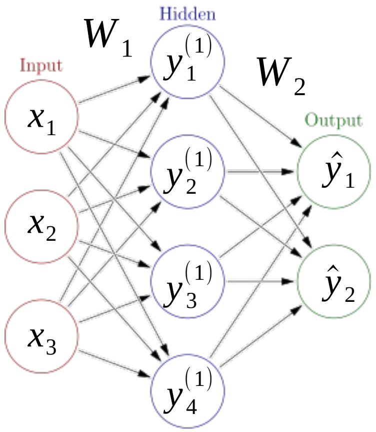
Gaussian process: Bayesian DNN connection

$$K(x,x') = \sqrt{\frac{1}{K}}\sigma(W_1 x + b)^T \sqrt{\frac{1}{K}}\sigma(W_1 x' + b) \sim y^{(1)}(W_1, x, b)^T y^{(1)}(W_1, x', b)$$

$$W_1 \sim K \times Q; \;\; x \sim Q \times 1; \;\; y^{(1)} \sim K \times 1$$

with W_1 and b random variables as usual in a Bayesian approach

σ: ReLU, sigmoid, etc.



To find the connection between a GP and DNN's let's make use of the evidence:

$$P(Y/X) = \int P(Y/f)\, P(f/W_1, b, X)\, P(W_1)\, P(b)\, df\, dW_1\, db$$

$P(Y/f) \sim N(Y, f, \tau^{-1} I_D)$ Likelihood

$P(f/W_1, b, X) \sim GP(0, K)$

$$\int df\, P(Y/f)\, P(f/W_1, b, X) = N(Y; 0, (K(X,X) + \tau^{-1}) I_D) = \int N(Y; W_2 y^{(1)}, \tau^{-1} I_D)\, P(W_2)\, dW_2$$

auxiliar matrix variables $: W_2$

$f(X, W_1, W_2, b) = \hat{y} = W_2 \sigma(W_1 X + b)$

$P(Y/W_1, W_2, b, X)$ likelihood in parameter space

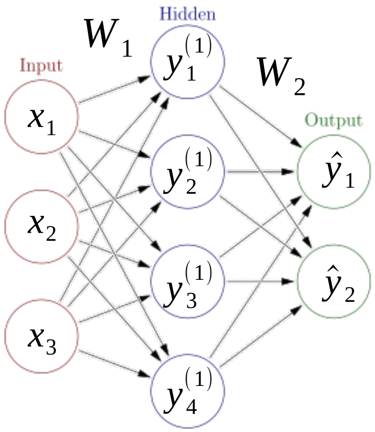
Gaussian process: Bayesian DNN connection

$$K(x,x')=\sqrt{\frac{1}{K}}\sigma(W_1x+b)^T\sqrt{\frac{1}{K}}\sigma(W_1x'+b)\sim y^{(1)}(W_1,x,b)^T y^{(1)}(W_1,x',b)$$

$$W_1\sim K\times Q;\;\; x\sim Q\times 1;\;\; y^{(1)}\sim K\times 1$$

with W_1 and b random variables as usual in a Bayesian approach

σ: ReLU, sigmoid, etc.



To find the connection between a GP and DNN's let's make use of the evidence:

$$P(Y/X)=\int P(Y/f)P(f/W_1,b,X)P(W_1)P(b)\,df\,dW_1\,db$$

$$=\int P(Y/W_1,W_2,b,X)P(W_1)P(W_2)P(b)\,dW_1\,dW_2\,db$$

Bayesian parametric representation of our 2 layer DNN !!

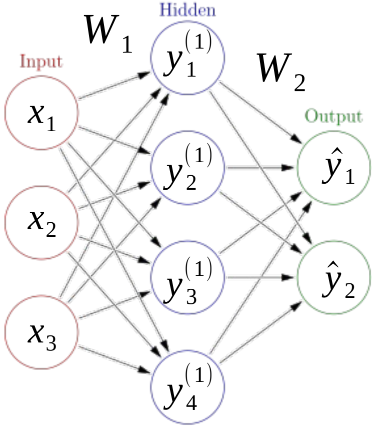
Gaussian process: Bayesian DNN connection

$$K(x,x')=\sqrt{\frac{1}{K}}\sigma(W_1x+b)^T\sqrt{\frac{1}{K}}\sigma(W_1x'+b)\sim y^{(1)}(W_1,x,b)^T y^{(1)}(W_1,x',b)$$

$$W_1\sim K\times Q;\ \ x\sim Q\times 1;\ \ y^{(1)}\sim K\times 1$$

with W_1 and b random variables as usual in a Bayesian approach

σ: ReLU, sigmoid, etc.



To find the connection between a GP and DNN's let's make use of the evidence:

$$P(Y/X)=\int P(Y/f)P(f/W_1,b,X)P(W_1)P(b)\,df\,dW_1\,db$$

$$=\int P(Y/W_1,W_2,b,X)P(W_1)P(W_2)P(b)\,dW_1\,dW_2\,db$$

Bayesian parametric representation of our 2 layer DNN !!

Conclusion: $K(x,x')$ right kernel to approximate DNN functions in a GP approach.

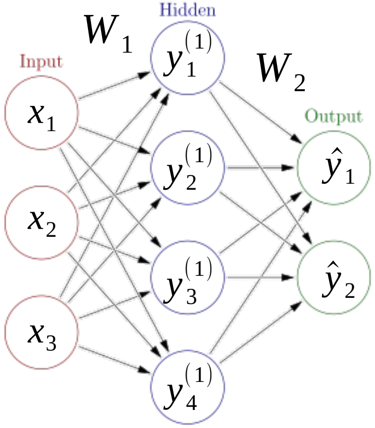
Gaussian process: Bayesian DNN connection

$$K(x,x')=\sqrt{\frac{1}{K}}\sigma(W_1x+b)^T\sqrt{\frac{1}{K}}\sigma(W_1x'+b)\sim y^{(1)}(W_1,x,b)^T y^{(1)}(W_1,x',b)$$

$$W_1\sim K\times Q;\;\; x\sim Q\times 1;\;\; y^{(1)}\sim K\times 1$$

with W_1 and b random variables as usual in a Bayesian approach

σ: ReLU, sigmoid, etc.



To find the connection between a GP and DNN's let's make use of the evidence:

$$P(Y/X)=\int P(Y/f)P(f/W_1,b,X)P(W_1)P(b)\,df\,dW_1\,db$$

$$=\int P(Y/W_1,W_2,b,X)P(W_1)P(W_2)P(b)\,dW_1\,dW_2\,db$$

Bayesian parametric representation of our 2 layer DNN !!

Non-linearities σ non-trivial Impact in C.I. of Bayes-DNN

Conclusion: $K(x,x')$ right kernel to approximate DNN functions in a GP approach.

Note:

For a generalization to deeper DNN and classification problems see ref.2 (Appendix)

But for a more accurate connection see “Deep neural networks as Gaussian processes”, Lee et al ‘2017.

3 - Dropout from a Bayesian point of view

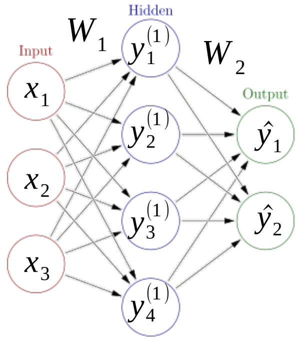
Variational Bayesian Inference and Dropout relationship

Applying Bayes theorem in a parametric space the predictive probability distribution of a DNN is given by:

$$P(\hat{y}/\hat{x},X,Y)=\int P(\hat{y}/\hat{x},W_1,W_2,b)\,P(W_1,W_2,b/X,Y)\,dW_1\,dW_2\,db$$

$$N(\hat{y};f=W_2\sigma(W_1\hat{x}+b),I)$$

σ: ReLU, sigmoid, etc.



Variational Bayesian Inference and Dropout relationship

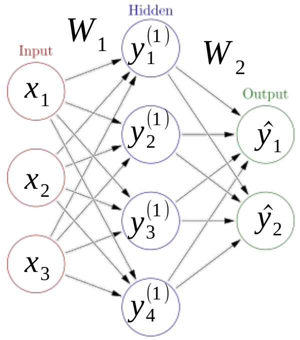
Applying Bayes theorem in a parametric space the predictive probability distribution of a DNN is given by:

$$P(\hat{y}/\hat{x},X,Y)=\int P(\hat{y}/\hat{x},W_1,W_2,b)\,P(W_1,W_2,b/X,Y)\,dW_1\,dW_2\,db$$

$$N(\hat{y};f=W_2\,\sigma(W_1\hat{x}+b),I)$$

Posterior: quite hard to compute !!

σ: ReLU, sigmoid, etc.



Variational Bayesian Inference and Dropout relationship

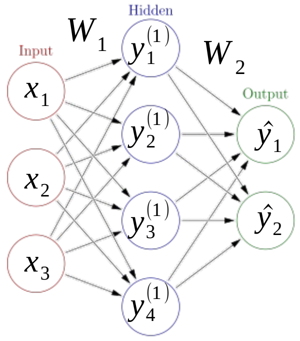
Applying Bayes theorem in a parametric space the predictive probability distribution of a DNN is given by:

$$P(\hat{y}/\hat{x},X,Y)=\int P(\hat{y}/\hat{x},W_1,W_2,b)\,P(W_1,W_2,b/X,Y)\,dW_1\,dW_2\,db$$

σ: ReLU, sigmoid, etc.

$$N(\hat{y};f=W_2\sigma(W_1\hat{x}+b),I)$$

Posterior: quite hard to compute !!



$$P(W_1,W_2,b/X,Y)=\frac{P(Y/W_1,W_2,b,X)\,P(W_1,W_2,b/X)}{P(Y/X)}$$

$$P(Y/X)=\int P(Y/W_1,W_2,b,X)\,P(W_1)\,P(W_2)\,P(b)\,dW_1\,dW_2\,db$$

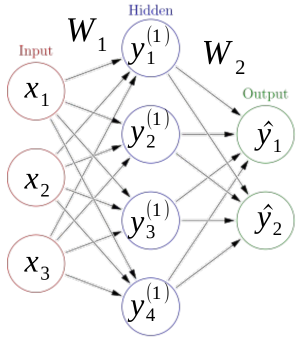
Intractable for large DNN !!!

Variational Bayesian Inference and Dropout relationship

Applying Bayes theorem in a parametric space the predictive probability distribution of a DNN is given by:

$$P(\hat{y}/\hat{x},X,Y)=\int P(\hat{y}/\hat{x},W_1,W_2,b)\,P(W_1,W_2,b/X,Y)\,dW_1\,dW_2\,db$$

σ: ReLU, sigmoid, etc.



$$N(\hat{y};f=W_2\sigma(W_1\hat{x}+b),I)$$

Posterior: quite hard to compute !!

$$P(W_1,W_2,b/X,Y)=\frac{P(Y/W_1,W_2,b,X)\,P(W_1,W_2,b/X)}{P(Y/X)}$$

$$P(Y/X)=\int P(Y/W_1,W_2,b,X)\,P(W_1)\,P(W_2)\,P(b)\,dW_1\,dW_2\,db$$

Intractable for large DNN !!!

Variational Inference: Approximation of the posterior by an anzat distribution.

Variational Bayesian Inference and Dropout relationship

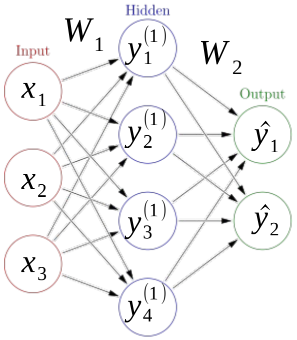
Applying Bayes theorem in a parametric space the predictive probability distribution of a DNN is given by:

$$P(\hat{y}/\hat{x},X,Y)=\int P(\hat{y}/\hat{x},W_1,W_2,b)\,P(W_1,W_2,b/X,Y)\,dW_1\,dW_2\,db$$

σ: ReLU, sigmoid, etc.

$$N(\hat{y};f=W_2\sigma(W_1\hat{x}+b),I)$$

Posterior: quite hard to compute !!



$$P(W_1,W_2 b/X,Y)\sim q_{\boldsymbol{M}}(W_1,W_2,b)=q_{M_1}(W_1)q_{M_2}(W_2)q_m(b)$$

Variational Bayesian Inference and Dropout relationship

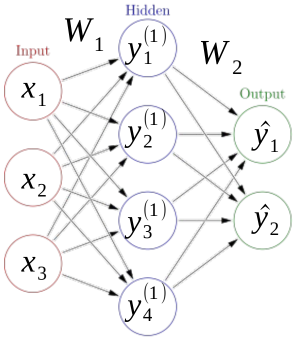
Applying Bayes theorem in a parametric space the predictive probability distribution of a DNN is given by:

$$P(\hat{y}/\hat{x},X,Y)=\int P(\hat{y}/\hat{x},W_1,W_2,b)\,P(W_1,W_2,b/X,Y)\,dW_1\,dW_2\,db$$

σ: ReLU, sigmoid, etc.

$$N(\hat{y};f=W_2\sigma(W_1\hat{x}+b),I)$$

Posterior: quite hard to compute !!



$$P(W_1,W_2 b/X,Y)\sim q_{\boldsymbol{M}}(W_1,W_2,b)=q_{M_1}(W_1)q_{M_2}(W_2)q_m(b)$$

$q_M(W)=\prod_\alpha q_{m_\alpha}(w_\alpha)$ with w_α/m_α the colums of W/M

$$q_{m_\alpha}(w_\alpha)=pN(m_\alpha,\theta^2 I)+(1-p)*N(0,\theta^2 I)$$

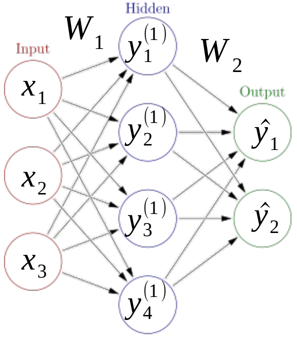
$$q(b)=N(m,\theta^2 I)$$

Variational Bayesian Inference and Dropout relationship

Applying Bayes theorem in a parametric space the predictive probability distribution of a DNN is given by:

$$P(\hat{y}/\hat{x}, X, Y) = \int P(\hat{y}/\hat{x}, W_1, W_2, b)\, P(W_1, W_2, b/X, Y)\, dW_1\, dW_2\, db$$

σ: ReLU, sigmoid, etc.



$$N(\hat{y}; f = W_2 \sigma(W_1 \hat{x} + b), I)$$

Posterior: quite hard to compute !!

$$P(W_1, W_2 b/X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1)\, q_{M_2}(W_2)\, q_m(b)$$

$q_M(W) = \prod_\alpha q_{m_\alpha}(w_\alpha)$ with w_α / m_α the colums of W/M

$$q_{m_\alpha}(w_\alpha) = p\, N(m_\alpha, \theta^2 I) + (1-p) * N(0, \theta^2 I)$$

$$q(b) = N(m, \theta^2 I)$$

Probability version of standard dropout approach !!

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

Kullback – Leibler divergence: $\omega \stackrel{\text{def}}{=} (W_1, W_2, b)$

$$KL(q_{\boldsymbol{M}}(\omega) | P(\omega / D)) = \int q_{\boldsymbol{M}}(\omega) \ln[P(D, \omega) / q_{\boldsymbol{M}}(\omega)] d\omega - \ln P(D) \quad \Rightarrow \quad ELBO(q_{\boldsymbol{M}}(\omega)) \leq \ln P(D)$$

≥ 0 $\qquad ELBO(q_{\boldsymbol{M}}(\omega))$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b/X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

Kullback – Leibler divergence: $\omega \stackrel{\text{def}}{=} (W_1, W_2, b)$

$$KL(q_{\boldsymbol{M}}(\omega) | P(\omega/D)) = \int q_{\boldsymbol{M}}(\omega) \ln[P(D,\omega)/q_{\boldsymbol{M}}(\omega)] d\omega - \ln P(D)$$

≥ 0 $\quad ELBO(q_{\boldsymbol{M}}(\omega))$

$\Rightarrow$ $ELBO(q_{\boldsymbol{M}}(\omega)) \leq \ln P(D)$

$$\text{Max}_{\boldsymbol{M}} ELBO(q_{\boldsymbol{M}}(\omega)) \Rightarrow q_{\vec{M}}(\omega) \rightarrow P(D/\omega)$$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$\text{Max}_{\boldsymbol{M}} ELBO(q_{\boldsymbol{M}}(W_1, W_2, b))$ ⇨ $q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D / W_1, W_2, b)$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b/X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}}\, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \quad \Rightarrow \quad q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

Likelihood

prior

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}}\, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \quad \Rightarrow \quad q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D / W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D / W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D / W_1, W_2, b)] = \int \ln P(D / W_1, W_2, b) q_{\boldsymbol{M}}(W_1, W_2, b) dW_1 dW_2 db$$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}} \, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \Rightarrow q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] = \int \ln P(D/W_1, W_2, b) q_{\boldsymbol{M}}(W_1, W_2, b) dW_1 dW_2 db$$

$$\ln P(D/W_1, W_2, b) = \ln N(Y; \hat{Y} = f(X, W_1, W_2, b), \tau^{-1} I_D) \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - \hat{y}_n\| \quad \text{with} \quad \hat{y}_n = f(x_n, W_1, W_2, b)$$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}} \, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \quad \Rightarrow \quad q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] = \int \ln P(D/W_1, W_2, b) q_{\boldsymbol{M}}(W_1, W_2, b) dW_1 dW_2 db$$

$$\ln P(D/W_1, W_2, b) = \ln N(Y; \hat{Y} = f(X, W_1, W_2, b), \tau^{-1} I_D) \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - \hat{y}_n\| \quad \text{with} \quad \hat{y}_n = f(x_n, W_1, W_2, b)$$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$\text{Max}_{\boldsymbol{M}} ELBO(q_{\boldsymbol{M}}(W_1, W_2, b))$ ⇒ $q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D / W_1, W_2, b)$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D / W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D / W_1, W_2, b)] = \frac{\tau}{2} \sum_{n=1}^{N} \int \|y_n - f(x_n, W_1, W_2, b)\| \times q_{\boldsymbol{M}}(W_1, W_2, b) \, dW_1 dW_2 db$$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$\text{Max}_{\boldsymbol{M}} \, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b))$ ➡ $q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] = \frac{\tau}{2} \sum_{n=1}^{N} \int \|y_n - f(x_n, W_1, W_2, b)\| \times q_{\boldsymbol{M}}(W_1, W_2, b) \, dW_1 dW_2 db$$

$\int \rightarrow \sum$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\mathrm{Max}_{\boldsymbol{M}}\, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \quad \Rightarrow \quad q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D / W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D / W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D / W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \sum_{\alpha=1}^{M} \|y_n - f(x_n, W_1^\alpha, W_2^\alpha, b^\alpha)\| \qquad W_1^\alpha, W_2^\alpha, b^\alpha \sim q_{\vec{M}}(W_1, W_2, b)$$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b/X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}}\, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \Rightarrow q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \sum_{\alpha=1}^{M} \|y_n - f(x_n, W_1^{\alpha}, W_2^{\alpha}, b^{\alpha})\| \qquad W_1^{\alpha}, W_2^{\alpha}, b^{\alpha} \sim q_{\vec{M}}(W_1, W_2, b)$$

$$\text{for } N \gg 1, \ \sum_n \sum_{\alpha} \rightarrow \sum_n$$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b/X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$\text{Max}_{\boldsymbol{M}} ELBO(q_{\boldsymbol{M}}(W_1, W_2, b))$ ➡ $q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - f(x_n, W_1^n, W_2^n, b^n)\|$$

$W_1^n, W_2^n, b^n \sim q_{\vec{M}}(W_1, W_2, b)$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b/X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$\text{Max}_{\boldsymbol{M}}\, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b))$ ➡ $q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - f(x_n, W_1^n, W_2^n, b^n)\| \qquad W_1^n, W_2^n, b^n \sim q_{\vec{M}}(W_1, W_2, b)$$

A bit of reparameterization:

$W_1 = diag(z_1)(M_1 + \theta \epsilon_1) + (Id - diag(z_1))\, \theta \epsilon_1$ (same for W_2)

$b = m + \theta \epsilon$ with $z_i \sim Bernoulli(p_i)$ and $\epsilon_i \sim N(0, I)$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b/X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$\text{Max}_{\boldsymbol{M}}\, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b))$ ⇒ $q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - f(x_n, W_1^n, W_2^n, b^n)\|$$

$W_1^n, W_2^n, b^n \sim q_{\vec{M}}(W_1, W_2, b)$

A bit of reparameterization:

$W_1 = diag(z_1)(M_1 + \theta \epsilon_1) + (Id - diag(z_1))\,\theta \epsilon_1$ (same for W_2)

$b = m + \theta \epsilon$ with $z_i \sim Bernoulli(p_i)$ and $\epsilon_i \sim N(0, I)$

$\theta \rightarrow 0$

$W_{1,2} \simeq diag(z_{1,2}) M_{1,2}$

$b \simeq m$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b/X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}}\, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \quad \Rightarrow \quad q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - f(x_n, W_1^n, W_2^n, b^n)\| \qquad W_1^n, W_2^n, b^n \sim q_{\vec{M}}(W_1, W_2, b)$$

A bit of reparameterization:

$W_1 = diag(z_1)(M_1 + \theta \epsilon_1) + (Id - diag(z_1))\theta \epsilon_1$ (same for W_2)

$b = m + \theta \epsilon$ with $z_i \sim Bernoulli(p_i)$ and $\epsilon_i \sim N(0, I)$

$\Rightarrow$ $f(x_n, W_1^n, W_2^n, b^n) = \hat{y}_n = M^{(2)} diag(z^{(2)}) \sigma(M^{(1)} diag(z^{(1)}) x + m)$

$\theta \rightarrow 0$

$W_{1,2} \simeq diag(z_{1,2}) M_{1,2}$

$b \simeq m$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b/X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}} ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \quad \Rightarrow \quad q_{\vec{M}}(W_1, W_2, b) \to P(D/W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$\hat{y}_n = M^{(2)} diag(z^{(2)}) \sigma(M^{(1)} diag(z^{(1)}) x + m)$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - f(x_n, W_1^n, W_2^n, b^n)\| \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - \hat{y}_n\|$$

$$\theta \to 0$$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b/X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}} \, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \Rightarrow q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$\hat{y}_n = M^{(2)} diag(z^{(2)}) \sigma(M^{(1)} diag(z^{(1)}) x + m)$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - f(x_n, W_1^n, W_2^n, b^n)\| \underset{\theta \rightarrow 0}{\sim} \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - \hat{y}_n\|$$

First term of dropout MLE loss

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\mathrm{Max}_{\boldsymbol{M}}\, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \quad \Rightarrow \quad q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D / W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D / W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$\hat{y}_n = M^{(2)} diag(z^{(2)}) \sigma(M^{(1)} diag(z^{(1)}) x + m)$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D / W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - f(x_n, W_1^n, W_2^n, b^n)\| \underset{\theta \rightarrow 0}{\sim} \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - \hat{y}_n\|$$

First term of dropout MLE loss

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}}\, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \Rightarrow q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D/W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$\hat{y}_n = M^{(2)} diag(z^{(2)}) \sigma(M^{(1)} diag(z^{(1)}) x + m)$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D/W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - f(x_n, W_1^n, W_2^n, b^n)\| \underset{\theta \rightarrow 0}{\sim} \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - \hat{y}_n\|$$

First term of dropout MLE loss

$$KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b)) \underset{\theta \rightarrow 0}{\sim} -\frac{p_1}{2}\|M_1\| - \frac{p_2}{2}\|M_2\| - \frac{1}{2}\|m\|$$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}} ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \quad \Rightarrow \quad q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D / W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) = E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D / W_1, W_2, b)] - KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b))$$

$$\hat{y}_n = M^{(2)} diag(z^{(2)}) \sigma(M^{(1)} diag(z^{(1)}) x + m)$$

$$E_{W_1, W_2, b \sim q_{\boldsymbol{M}}(W_1, W_2, b)}[\ln P(D / W_1, W_2, b)] \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - f(x_n, W_1^n, W_2^n, b^n)\| \underset{\theta \rightarrow 0}{\sim} \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - \hat{y}_n\|$$

First term of dropout MLE loss

$$KL(q_{\boldsymbol{M}}(W_1, W_2, b) | P(W_1, W_2, b)) \underset{\theta \rightarrow 0}{\sim} -\frac{p_1}{2} \|M_1\| - \frac{p_2}{2} \|M_2\| - \frac{1}{2} \|m\|$$

Regularization terms of dropout MLE loss

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$\mathrm{Max}_{\boldsymbol{M}}\, ELBO(q_{\boldsymbol{M}}(W_1, W_2, b))$ ⇒ $q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D / W_1, W_2, b)$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

Loss function of standard dropout!!!

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - \hat{y}_n\| - \frac{p_1}{2} \|M_1\| - \frac{p_2}{2} \|M_2\| - \frac{1}{2} \|m\|$$

$$\hat{y}_n = M^{(2)} diag(z^{(2)}) \sigma(M^{(1)} diag(z^{(1)}) x + m)$$

Variational Bayesian Inference and Dropout relationship

Let's find the $\boldsymbol{M}$ optimal parameters: $P(W_1, W_2 b / X, Y) \sim q_{\boldsymbol{M}}(W_1, W_2, b) = q_{M_1}(W_1) q_{M_2}(W_2) q_m(b)$

$$\text{Max}_{\boldsymbol{M}} ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \quad \Rightarrow \quad q_{\vec{M}}(W_1, W_2, b) \rightarrow P(D / W_1, W_2, b)$$

From ELBO definition: (spoiler: $ELBO \sim L_{dropout} = 1/(2N) \sum_{n=1}^{N} \|y_n - \hat{y}_n\| + \lambda_1 \|W_1\| + \lambda_2 \|W_2\| + \lambda_3 \|b\|$)

Loss function of standard dropout!!!

$$ELBO(q_{\boldsymbol{M}}(W_1, W_2, b)) \sim \frac{\tau}{2} \sum_{n=1}^{N} \|y_n - \hat{y}_n\| - \frac{p_1}{2} \|M_1\| - \frac{p_2}{2} \|M_2\| - \frac{1}{2} \|m\|$$

$$\hat{y}_n = M^{(2)} diag(z^{(2)}) \sigma(M^{(1)} diag(z^{(1)}) x + m)$$

We will maximize the ELBO through standard MLE methods (Gradiend descent, etc)

Inference: let’s compute the predictions and variances

$$E(y^*) = \int y^* P(y^*/x^*, X, Y)\, dy^* = \int y^* P(y^*/x^*, W_1, W_2, b) * P(W_1, W_2, b/X, Y)\, dW_1\, dW_2\, db\, dy^*$$

Inference: let's compute the predictions and variances

$$E(y^*) = \int y^* P(y^*/x^*, X, Y)\, dy^* = \int y^* P(y^*/x^*, W_1, W_2, b) * P(W_1, W_2, b/X, Y)\, dW_1\, dW_2\, db\, dy^*$$

$$P(W_1, W_2, b/X, Y) \rightarrow q_M(W_1, W_2, b)$$

Inference: let's compute the predictions and variances

$$E(y^*) = \int y^* P(y^*/x^*, X, Y) dy^* = \int y^* P(y^*/x^*, W_1, W_2, b) * P(W_1, W_2, b/X, Y) dW_1 dW_2 db dy^*$$

$$P(W_1, W_2, b/X, Y) \rightarrow q_M(W_1, W_2, b)$$

$$E(y^*) \simeq \frac{1}{T} \sum_{t=1}^{T} \int y^* P(y^*/x^*, W_{1,t}, W_{2,t}, b_t) * d y^*$$

$$W_1 = diag(z_1) M_1$$
$$W_2 = diag(z_2) M_2$$
$$b = m$$

Inference: let's compute the predictions and variances

$$E(y^*) = \int y^* P(y^*/x^*, X, Y) dy^* = \int y^* P(y^*/x^*, W_1, W_2, b) * P(W_1, W_2, b/X, Y) dW_1 dW_2 db\, dy^*$$

$$P(W_1, W_2, b/X, Y) \rightarrow q_M(W_1, W_2, b)$$

$$\hat{y}^* = M^{(2)} diag(z^{(2)}) \sigma(M^{(1)} diag(z^{(1)}) x^* + m)$$

$$E(y^*) \simeq \frac{1}{T} \sum_{t=1}^{T} \int y^* P(y^*/x^*, W_{1,t}, W_{2,t}, b_t) * d\,y^* = \frac{1}{T} \sum_{t=1}^{T} \int y^* N(y^*, \hat{y}^*(x^*, z_{1,t}, z_{2,t})) * d\,y^*$$

$$W_1 = diag(z_1) M_1$$
$$W_2 = diag(z_2) M_2$$
$$b = m$$

Inference: let's compute the predictions and variances

$$E(y^*) = \int y^* P(y^*/x^*, X, Y)\, dy^* = \int y^* P(y^*/x^*, W_1, W_2, b) * P(W_1, W_2, b/X, Y)\, dW_1\, dW_2\, db\, dy^*$$

$$P(W_1, W_2, b/X, Y) \rightarrow q_M(W_1, W_2, b)$$

$$\hat{y}^* = M^{(2)} diag(z^{(2)})\, \sigma(M^{(1)} diag(z^{(1)}) x^* + m)$$

$$E(y^*) \simeq \frac{1}{T} \sum_{t=1}^{T} \int y^* P(y^*/x^*, W_{1,t}, W_{2,t}, b_t) * d\, y^* = \frac{1}{T} \sum_{t=1}^{T} \int y^* N(y^*, \hat{y}^*(x^*, z_{1,t}, z_{2,t})) * d\, y^*$$

$$= \hat{y}^*(x^*, z_{1,t}, z_{2,t})$$

$$W_1 = diag(z_1) M_1$$
$$W_2 = diag(z_2) M_2$$
$$b = m$$

Inference: let’s compute the predictions and variances

$$E(y^*) = \int y^* P(y^*/x^*, X, Y)\,dy^* = \int y^* P(y^*/x^*, W_1, W_2, b) * P(W_1, W_2, b/X, Y)\,dW_1\,dW_2\,db\,dy^*$$

$$P(W_1, W_2, b/X, Y) \rightarrow q_M(W_1, W_2, b)$$

$$\hat{y}^* = M^{(2)} diag(z^{(2)})\,\sigma(M^{(1)} diag(z^{(1)}) x^* + m)$$

$$E(y^*) \simeq \frac{1}{T}\sum_{t=1}^{T} \int y^* P(y^*/x^*, W_{1,t}, W_{2,t}, b_t) * d\,y^* = \frac{1}{T}\sum_{t=1}^{T} \int y^* N(y^*, \hat{y}^*(x^*, z_{1,t}, z_{2,t})) * d\,y^* = \frac{1}{T}\sum_{t=1}^{T} \hat{y}^*(x^*, z_{1,t}, z_{2,t})$$

$$= \hat{y}^*(x^*, z_{1,t}, z_{2,t})$$

$$W_1 = diag(z_1) M_1$$
$$W_2 = diag(z_2) M_2$$
$$b = m$$

Inference: let's compute the predictions and variances

$$E(y^*) = \int y^* P(y^*/x^*, X, Y) dy^* = \int y^* P(y^*/x^*, W_1, W_2, b) * P(W_1, W_2, b/X, Y) dW_1 dW_2 db\, dy^*$$

$$P(W_1, W_2, b/X, Y) \rightarrow q_M(W_1, W_2, b)$$

$$E(y^*) \simeq \frac{1}{T} \sum_{t=1}^{T} \hat{y}^*(x^*, z_{1,t}, z_{2,t})$$

Inference: let's compute the predictions and variances

$$E(y^*) = \int y^* P(y^*/x^*, X, Y)\,dy^* = \int y^* P(y^*/x^*, W_1, W_2, b) * P(W_1, W_2, b/X, Y)\,dW_1\,dW_2\,db\,dy^*$$

$$P(W_1, W_2, b/X, Y) \rightarrow q_M(W_1, W_2, b)$$

$$E(y^*) \simeq \frac{1}{T}\sum_{t=1}^{T} \hat{y}^*(x^*, z_{1,t}, z_{2,t})$$

Dropout interpretation: Ensemble model

$\tilde{y}_1$	$\tilde{y}_2$	$\tilde{y}_3$	
1	1	1	Model 1
1	1	0	Model 2
	⋮		
0	0	0	Model 2^N

In agreement with our initial interpretation of ensemble model or model averaging !!

Inference: let’s compute the predictions and variances

MC dropout method

$$E_{q_M(y^*/x^*)}(y^*) \simeq \frac{1}{T}\sum_{t=1}^{T} \hat{y}^*(x^*, z_1^t, z_2^t, \ldots)$$ Mean

$$Var_{q_M(y^*/x^*)}(y^*) \simeq \tau^{-1} I_D + \frac{1}{T}\sum_{t=1}^{T} \hat{y}^*(x^*, z_1^t, z_2^t, \ldots)^T \hat{y}^*(x^*, z_1^t, z_2^t, \ldots) - E_{q_M(y^*/x^*)}(y^*)^T E_{q_M(y^*/x^*)}(y^*)$$ Variance

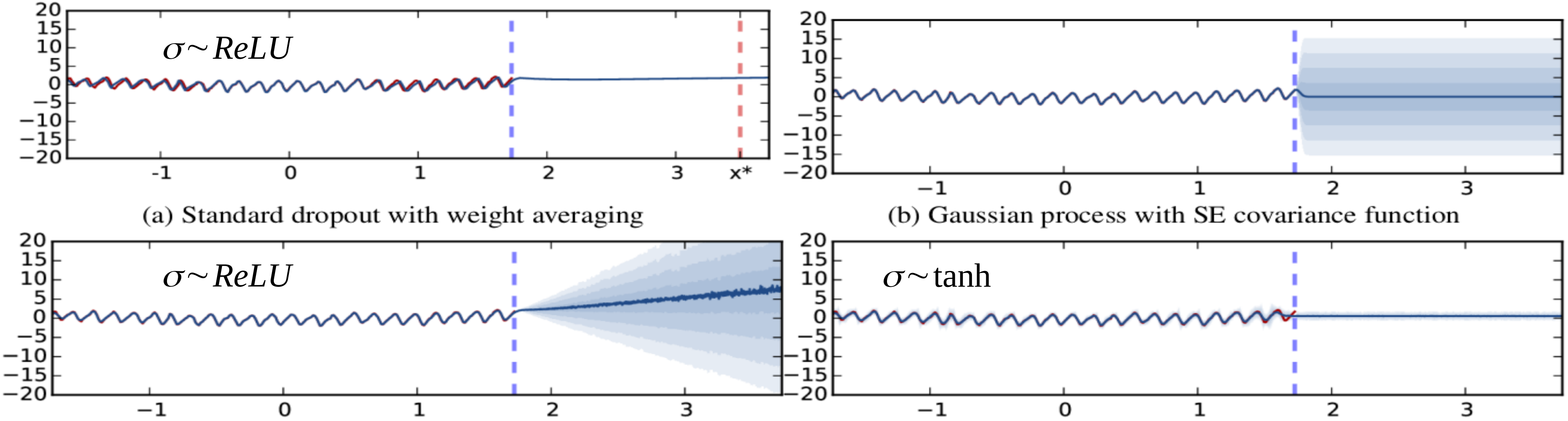
$$\hat{y}^*(x^*, z_1, z_2, \ldots) = (M_L\, diag(z_L))\, \sigma(\ldots(M_2\, diag(z_2))\, \sigma((M_1\, diag(z_1))\, x^* + m_1))$$

$$z_1, z_2 \sim Bern(p_1), Bern(p_2)$$

4 - Results

Regression

(a),(c) and (d): DNN with 4 layers and 1024 hidden units – p ~0.2



(a) Standard dropout with weight averaging

(b) Gaussian process with SE covariance function

(c) MC dropout with ReLU non-linearities

(d) MC dropout with TanH non-linearities

Figure 2. **Predictive mean and uncertainties on the Mauna Loa CO_2 concentrations dataset, for various models.** In red is the observed function (left of the dashed blue line); in blue is the predictive mean plus/minus two standard deviations (8 for fig. 2d). Different shades of blue represent half a standard deviation. Marked with a dashed red line is a point far away from the data: standard dropout confidently predicts an insensible value for the point; the other models predict insensible values as well but with the additional information that the models are uncertain about their predictions.

Regression

(a),(c) and (d): DNN with 4 layers and 1024 hidden units – p ~0.2

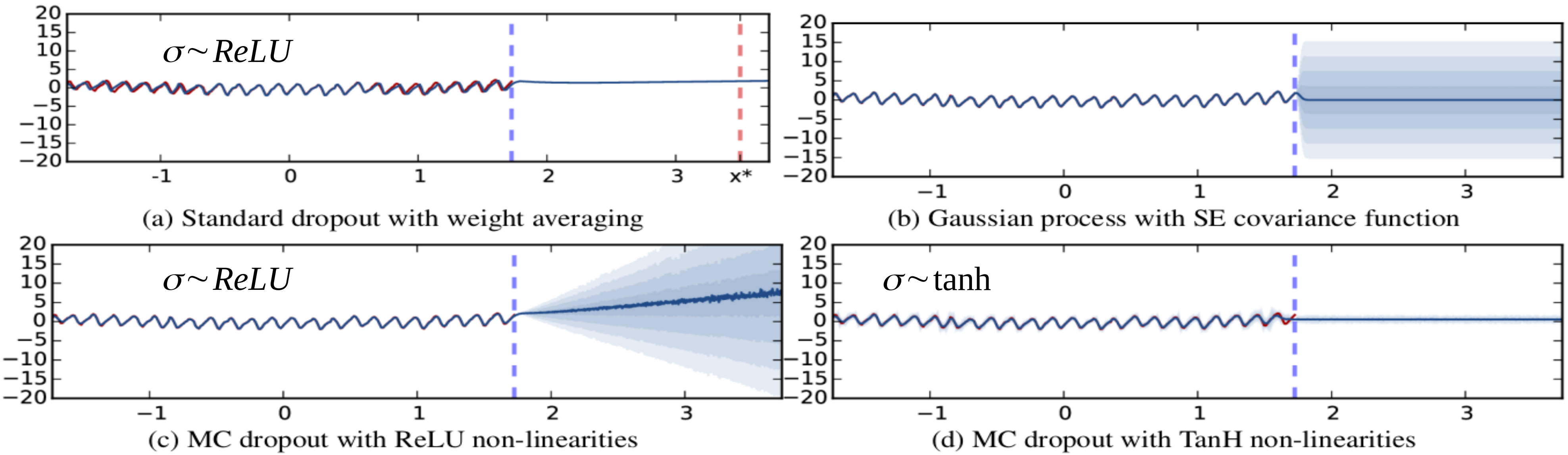
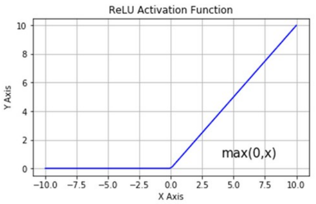
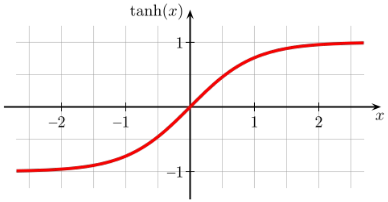


Figure 2. **Predictive mean and uncertainties on the Mauna Loa CO_2 concentrations dataset, for various models.** In red is the observed function (left of the dashed blue line); in blue is the predictive mean plus/minus two standard deviations (8 for fig. 2d). Different shades of blue represent half a standard deviation. Marked with a dashed red line is a point far away from the data: standard dropout confidently predicts an insensible value for the point; the other models predict insensible values as well but with the additional information that the models are uncertain about their predictions.



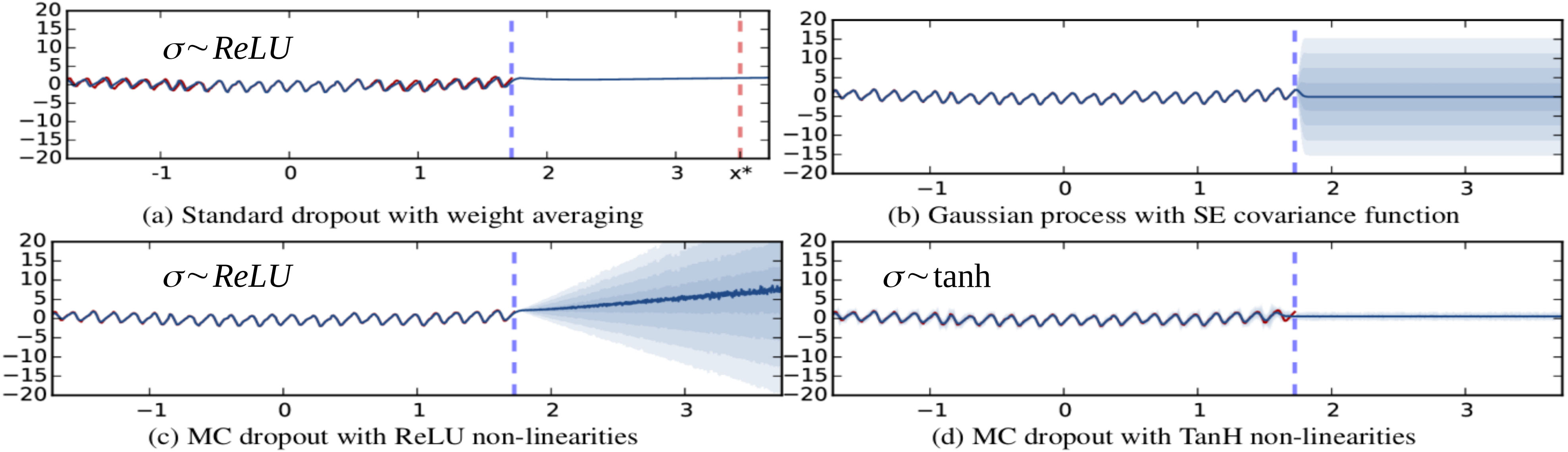
Variance ➡ $K(x,x') \simeq \sqrt{\frac{1}{K}}\sigma(W_1 x + b)^T \sqrt{\frac{1}{K}}\sigma(W_1 x' + b)$



Regression

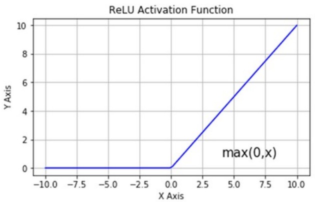
Q. Is it really realistic to change the non-linearity?

(a),(c) and (d): DNN with 4 layers and 1024 hidden units – p ~0.2

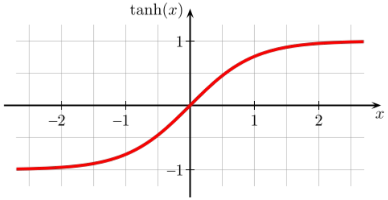


(a) Standard dropout with weight averaging
(b) Gaussian process with SE covariance function
(c) MC dropout with ReLU non-linearities
(d) MC dropout with TanH non-linearities

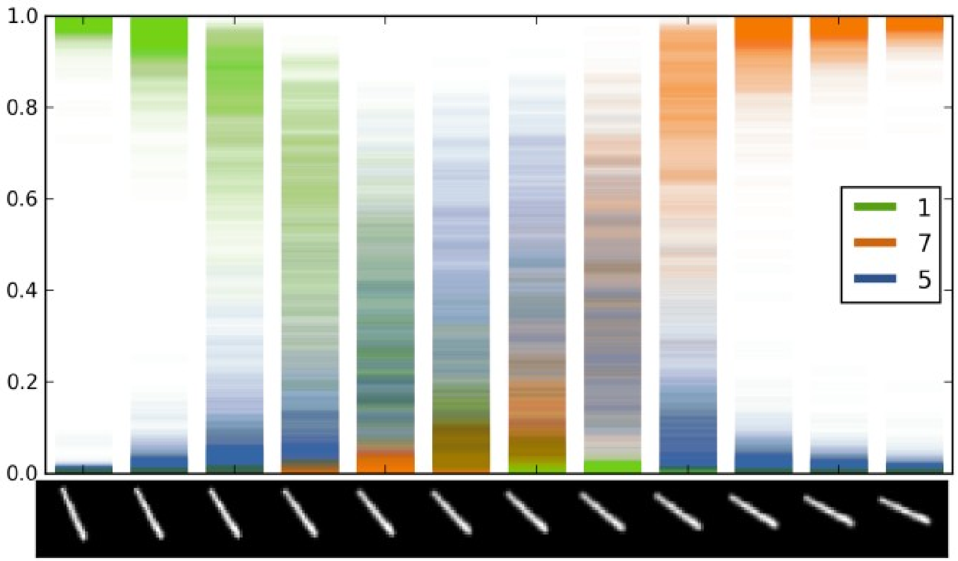
Figure 2. **Predictive mean and uncertainties on the Mauna Loa CO_2 concentrations dataset, for various models.** In red is the observed function (left of the dashed blue line); in blue is the predictive mean plus/minus two standard deviations (8 for fig. 2d). Different shades of blue represent half a standard deviation. Marked with a dashed red line is a point far away from the data: standard dropout confidently predicts an insensible value for the point; the other models predict insensible values as well but with the additional information that the models are uncertain about their predictions.



Variance → $$K(x,x') \simeq \sqrt{\frac{1}{K}}\sigma(\mathbf{W_1 x}+b)^T \sqrt{\frac{1}{K}}\sigma(\mathbf{W_1 x'}+b)$$



LeNet CNN on MNIST with dropout before last fully connected layer (p~0.5)

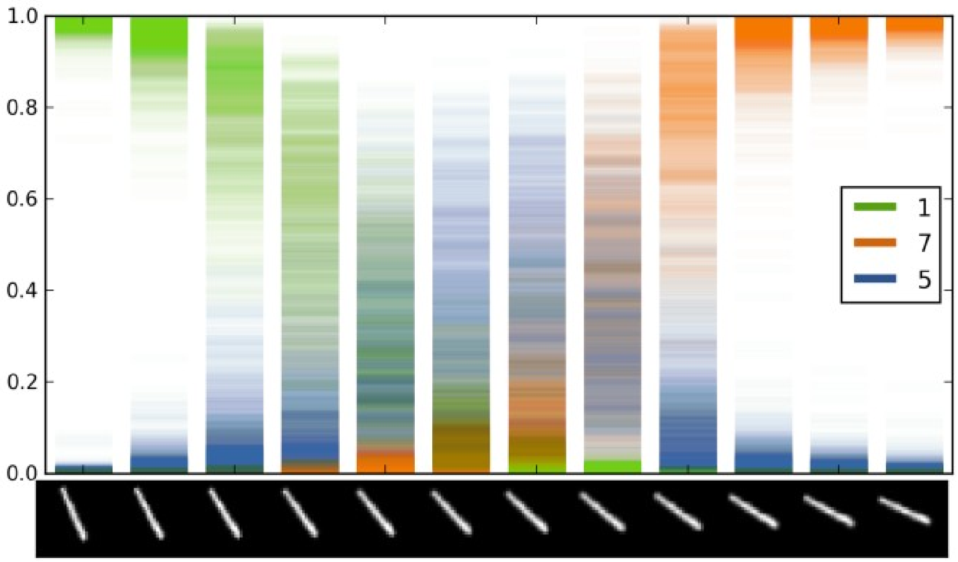


$$E_{q_M(y^*/x^*)}(y^*) \simeq \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t^*(x^*, z_1^t, z_2^t, \ldots)$$ T = 100

$$\hat{y}_t^* = (y_1, y_2, y_3, \ldots, y_{10})_t$$

Classification

LeNet CNN on MNIST with dropout before last fully connected layer (p~0.5)



$$E_{q_M(y^*/x^*)}(y^*) \simeq \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t^*(x^*, z_1^t, z_2^t, \ldots) \quad T = 100$$

$$\hat{y}_t^* = (y_1, y_2, y_3, \ldots, y_{10})_t$$

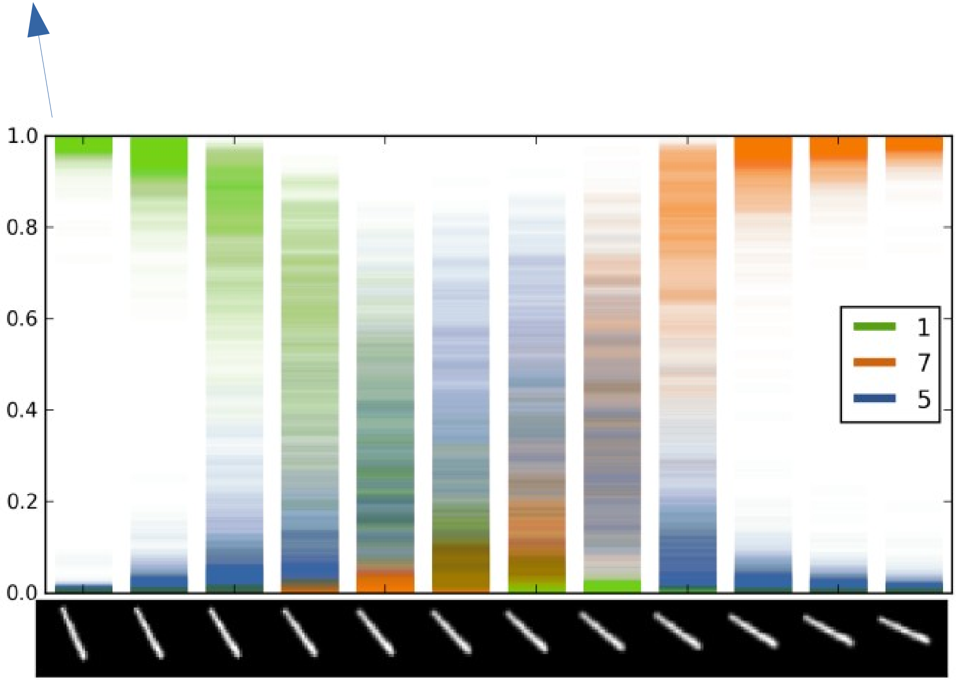
For every t plot largest three probs. of $(y_1, y_2, y_3, \ldots, y_{10})_t$

Classification

LeNet CNN on MNIST with dropout before last fully connected layer (p~0.5)

$E(\hat{y}_1)\sim 1,\ E(\hat{y}_5)\sim 0,\ E(\hat{y}_7)\sim 0$

$Var(\hat{y}_1)\sim 0,\ Var(\hat{y}_5)\sim 0,\ Var(\hat{y}_7)\sim 0$



$$E_{q_M(y^*/x^*)}(y^*)\simeq\frac{1}{T}\sum_{t=1}^{T}\hat{y}_t^*(x^*,z_1^t,z_2^t,\ldots)\quad T = 100$$

$$\hat{y}_t^*=(y_1,y_2,y_3,\ldots,y_{10})_t$$

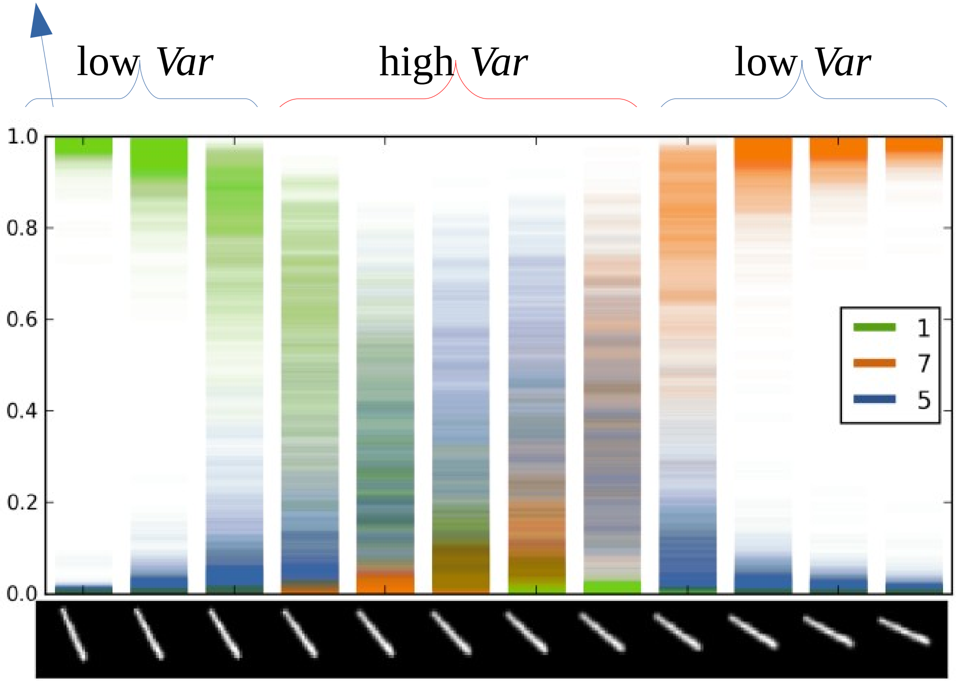
For every t plot largest three probs. of $(y_1,y_2,y_3,\ldots,y_{10})_t$

Classification

LeNet CNN on MNIST with dropout before last fully connected layer (p~0.5)

$E(\hat{y}_1) \sim 1,\ E(\hat{y}_5) \sim 0,\ E(\hat{y}_7) \sim 0$

$Var(\hat{y}_1) \sim 0,\ Var(\hat{y}_5) \sim 0,\ Var(\hat{y}_7) \sim 0$



$$E_{q_M(y^*/x^*)}(y^*) \simeq \frac{1}{T} \sum_{t=1}^{T} \hat{y}_t^*(x^*, z_1^t, z_2^t, \ldots) \quad T = 100$$

$$\hat{y}_t^* = (y_1, y_2, y_3, \ldots, y_{10})_t$$

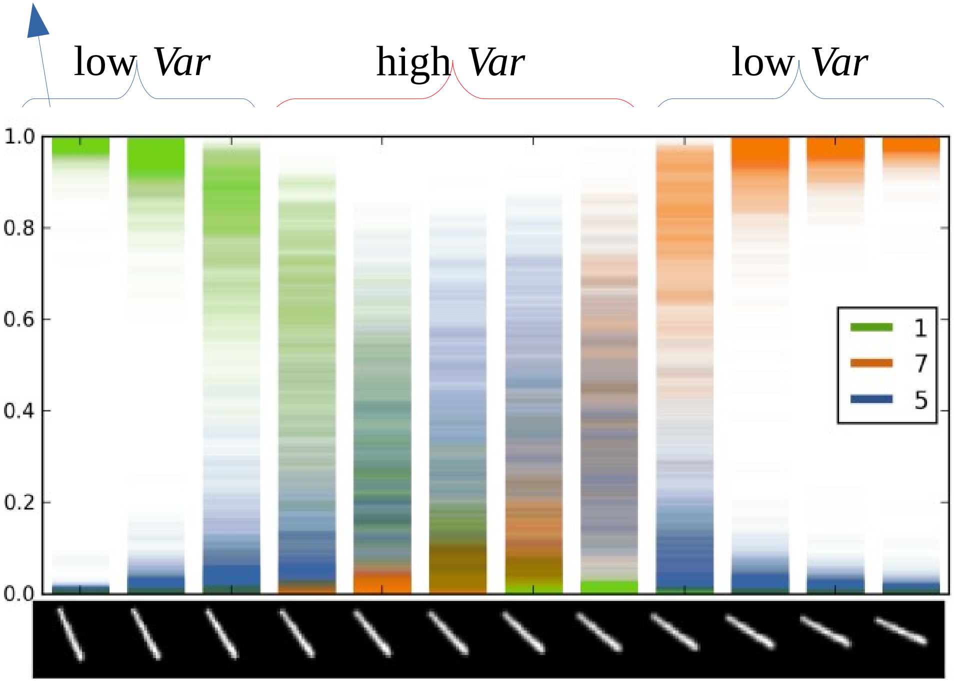
For every t plot largest three probs. of $(y_1, y_2, y_3, \ldots, y_{10})_t$

Classification

LeNet CNN on MNIST with dropout before last fully connected layer (p~0.5)

$E(\hat{y}_1) \sim 1,\ E(\hat{y}_5) \sim 0,\ E(\hat{y}_7) \sim 0$

$Var(\hat{y}_1) \sim 0,\ Var(\hat{y}_5) \sim 0,\ Var(\hat{y}_7) \sim 0$



Q. Is the model better calibrated in this way?

$$E_{q_M(y^*/x^*)}(y^*) \simeq \frac{1}{T} \sum_{t=1}^{T} \hat{y}_t^*(x^*, z_1^t, z_2^t, \ldots) \quad T = 100$$

$$\hat{y}_t^* = (y_1, y_2, y_3, \ldots, y_{10})_t$$

For every t plot largest three probs. of $(y_1, y_2, y_3, \ldots, y_{10})_t$

THANKS !!!