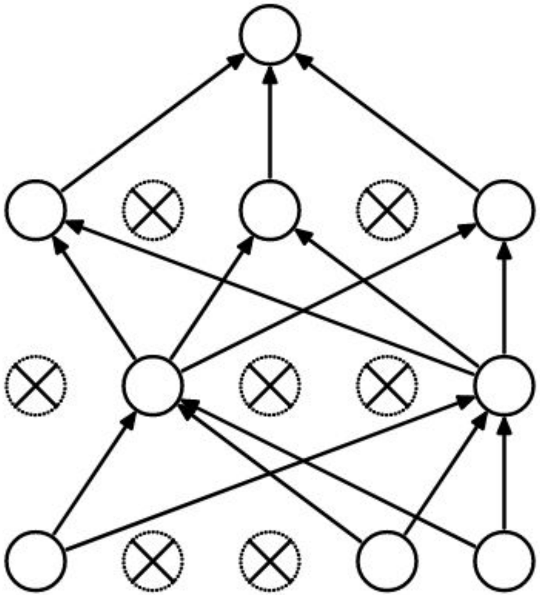
Dropout as a Structured Shrinkage Prior

Eric Nalisnick, Jose Miguel Hernandez-Lobato, Padhraic Smyth

Dropout: [Hinton et al '12]

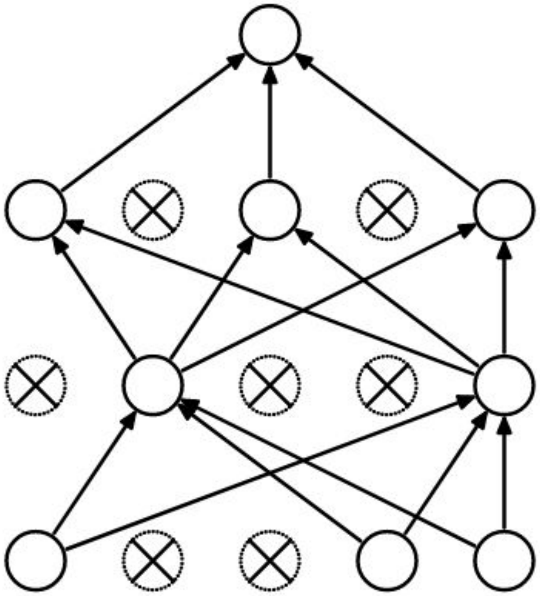


Train: randomly set weights to 0

Test:

- use all weights to predict

Dropout: [Gal & Ghahramani '16]

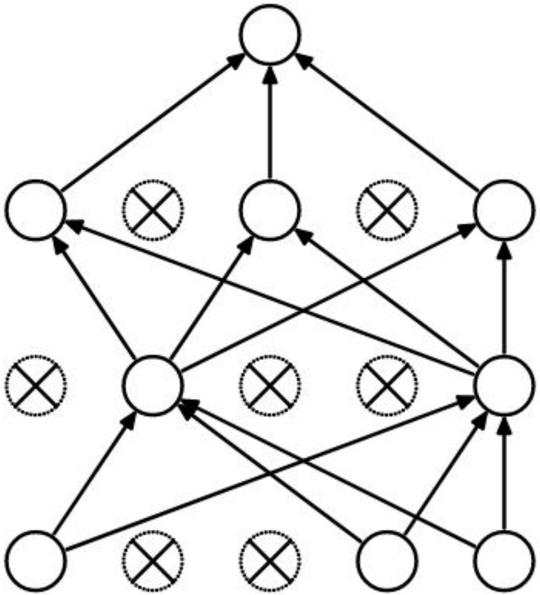


Train: randomly set weights to 0

Test:

- randomly set weights to 0
- use weights to predict
- average predictions

Dropout: [Gal & Ghahramani '16]



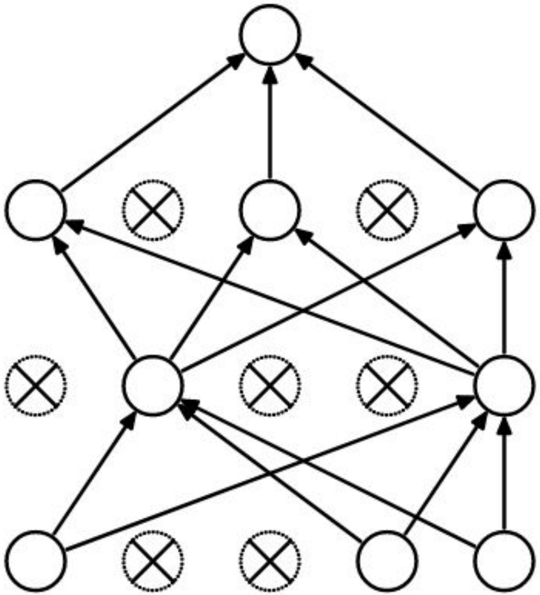
Train: randomly set weights to 0

Test:

- randomly set weights to 0
- use weights to predict
- average predictions

Marginalizing posterior by sampling

Dropout: [This paper]



Marginalizing posterior by sampling

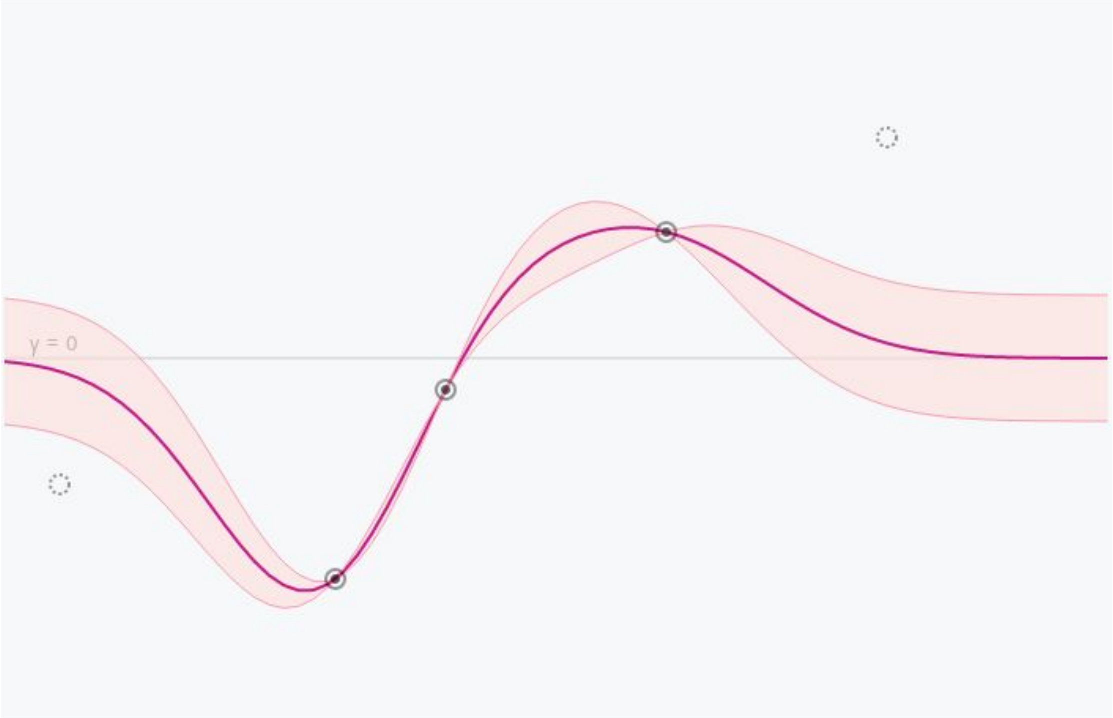
Multiplying bernoulli noise

Train: ~~randomly set weights to 0~~

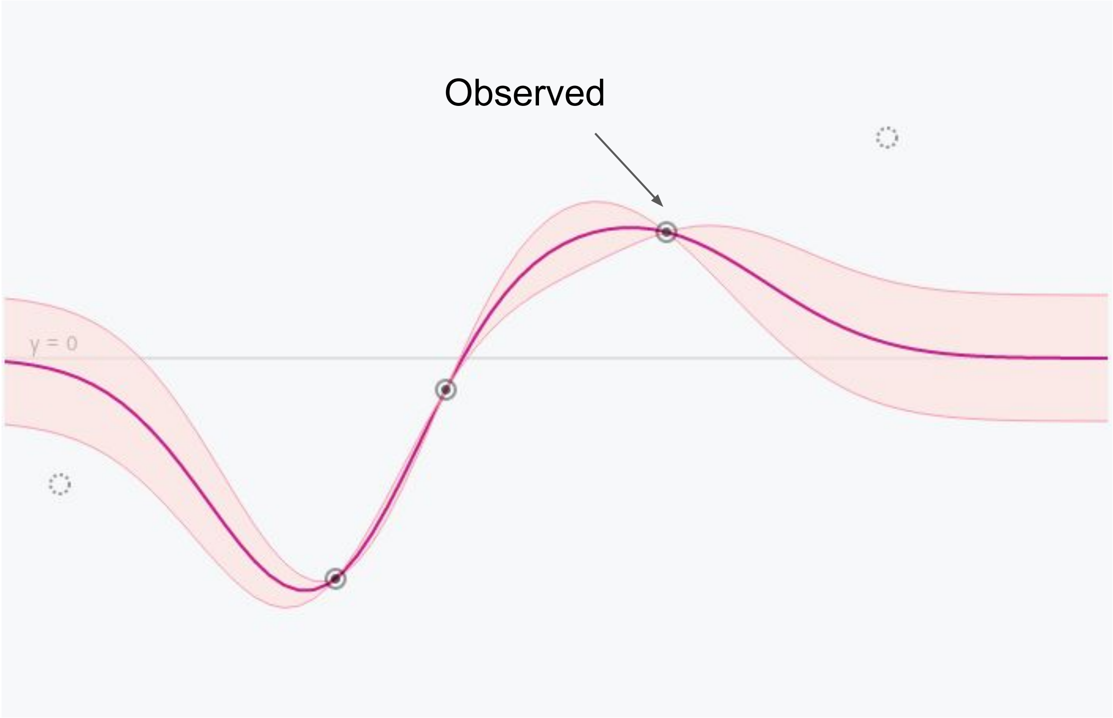
Test:

- randomly set weights to 0
- use weights to predict
- average predictions

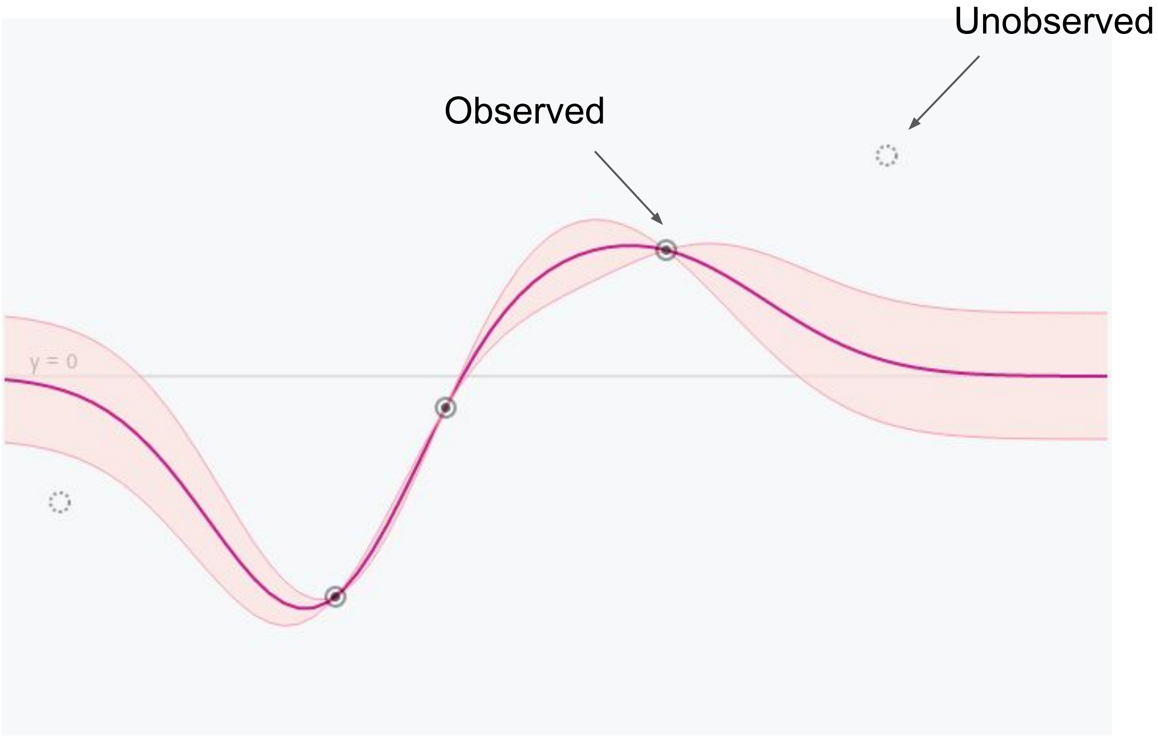
Regression



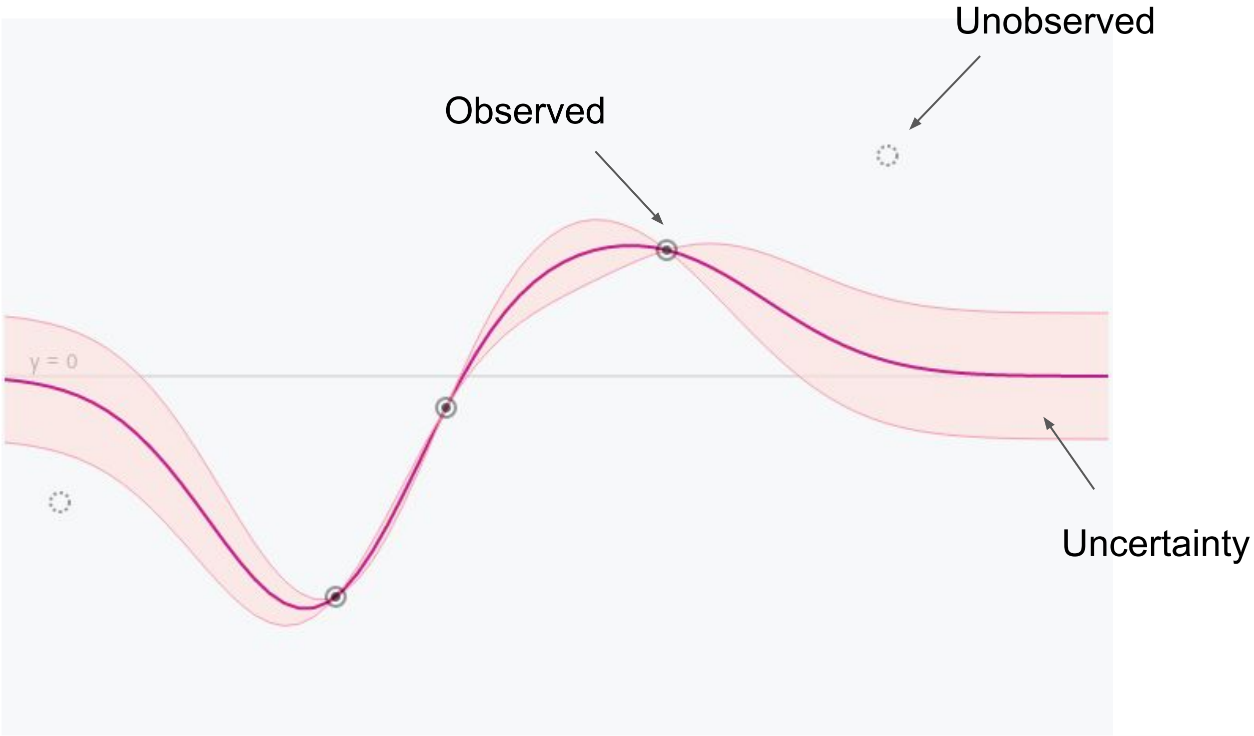
Regression



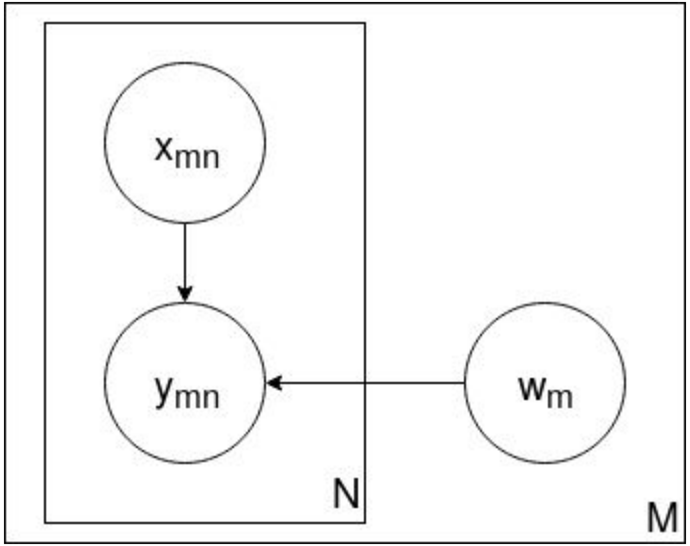
Regression



Regression

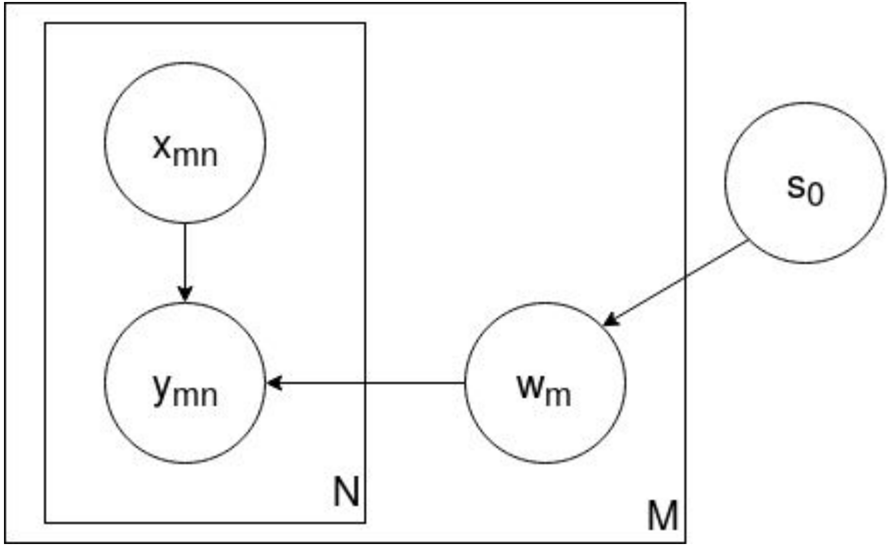


Regression: Maximum Likelihood



$y|w \sim N(f(w, x), s)$

Regression: Maximum a Posteriori

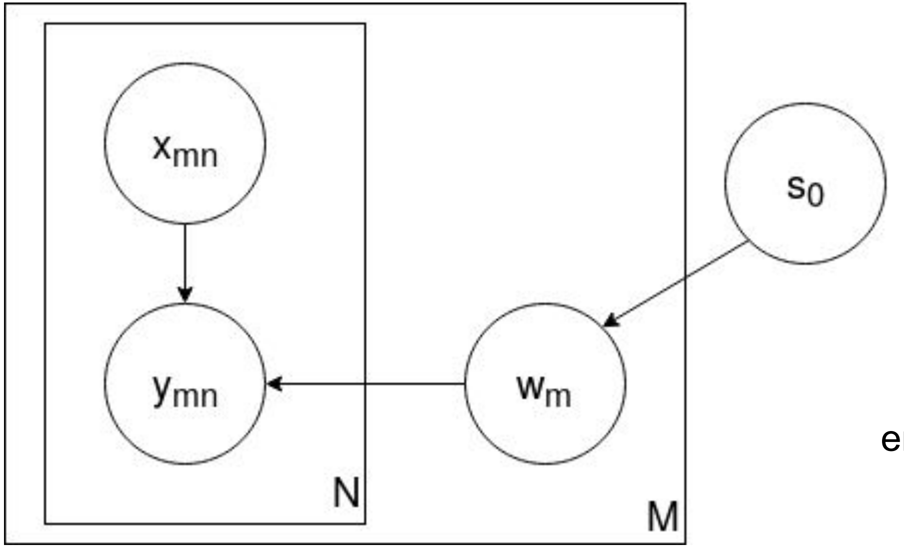


regularization $w \sim N(0, s_0)$

$y|w \sim N(f(w, x), s)$

$w|y \sim N(w_{map}, s_n)$

Regression: Full Bayesian



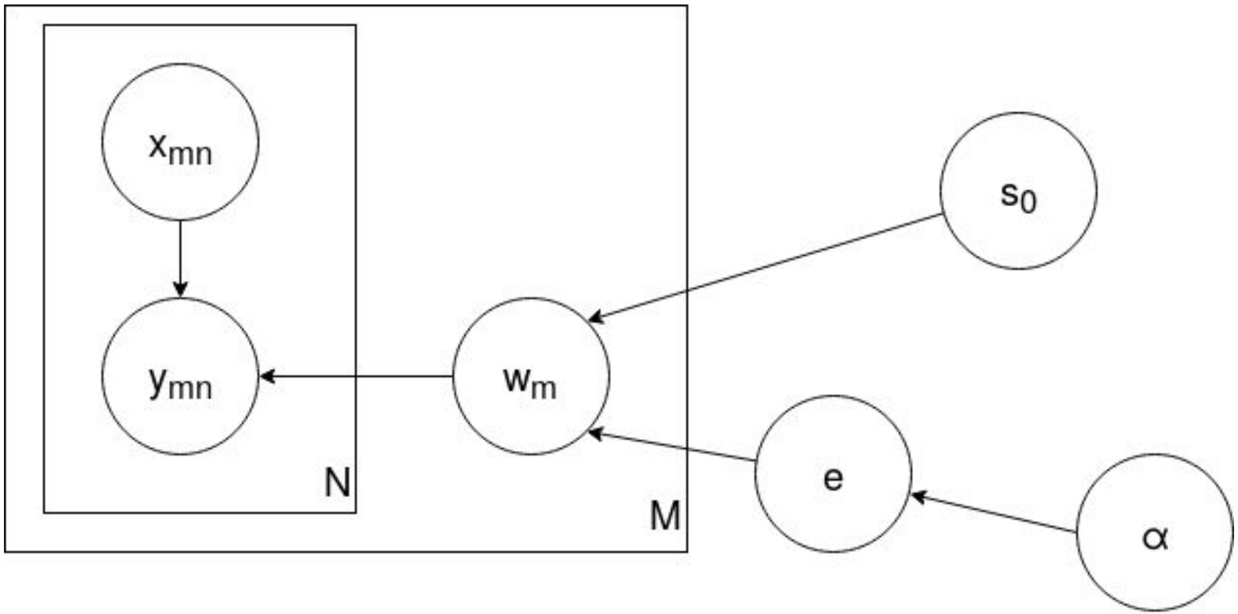
$w \sim N(0, s_0)$

$y|w \sim N(f(w, x), s)$

$w|y \sim N(w_{map}, s_n)$

ensembling $p(y|x) = \int_w p(w|y)\, p(y|w)\, dw$

I heard you like priors …



$e \sim \text{Exp}(\alpha)$

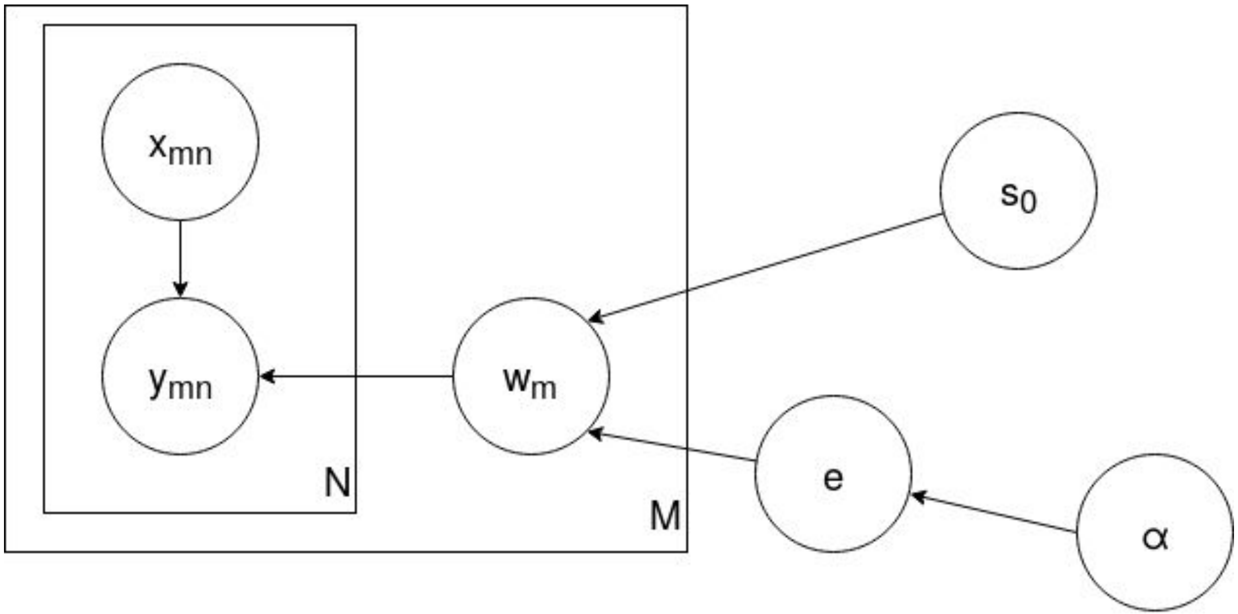
$w \sim N(0, es_0)$

$y|w \sim N(f(w, x), s)$

$w|y \sim N(w_{map}, s_n)$

I heard you like priors …

multiplicative noise $e \sim \text{Exp}(\alpha)$

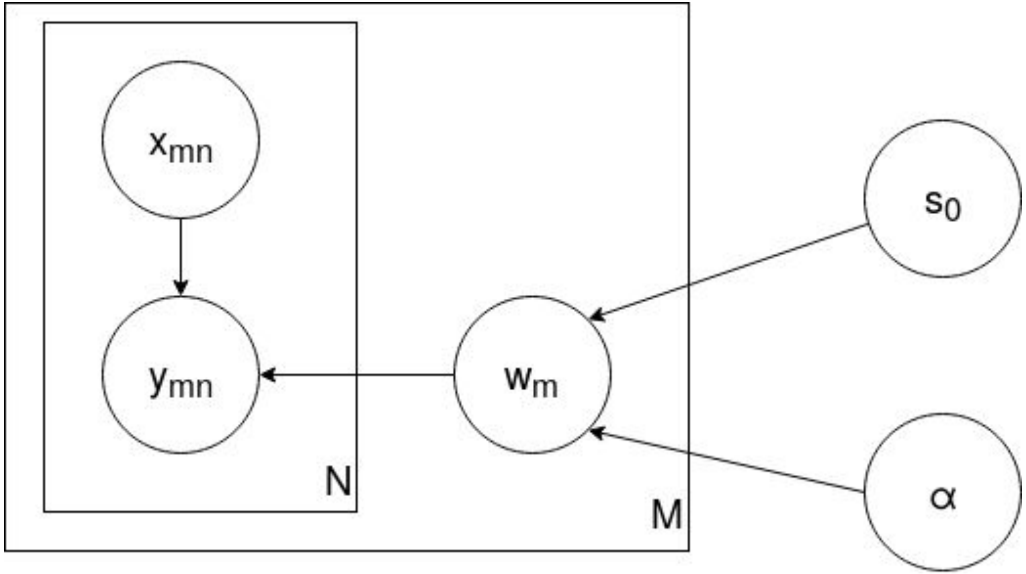


$w \sim N(0, es_0)$

$y|w \sim N(f(w, x), s)$

$w|y \sim N(w_{map}, s_n)$

Gaussian Scale Mixture, 'shrinkage prior'



Laplace distribution / L1 Reg.

w ~ $N(0, s_0)$ $Exp(\alpha)$

$y|w \sim N(f(w, x), s)$

$w|y \sim N(w_{map}, s_n)$

Dropout == GSM

Let's assume a **Gaussian prior on the NN weights**...

$$f_l(\mathbf{h}_{n,l-1}\underbrace{\mathbf{\Lambda}_l\mathbf{W}_l}_{\textbf{Definition of a Gaussian Scale Mixture}})$$

↓ SWITCH TO HIERARCHICAL PARAMETRIZATION ↓

$$f_l(\mathbf{h}_{n,l-1}\mathbf{W}_l)$$

$$w_{i,j} \sim \mathrm{N}(0, \lambda_{i,i}^2\sigma_0^2)$$

A Generalization of Dropout

Noise Model $p(\xi)$	Variance Prior $p(\xi^2)$	Marginal Prior $p(w)$
Bernoulli	Bernoulli	Spike-and-Slab
Gaussian	χ^2	Gen. Hyperbolic
Rayleigh	Exponential	Laplace
Inverse Nakagami	Γ^{-1}	Student-t
Half-Cauchy	Unnamed	Horseshoe

Table 1. Noise Models and their Corresponding GSM Prior.

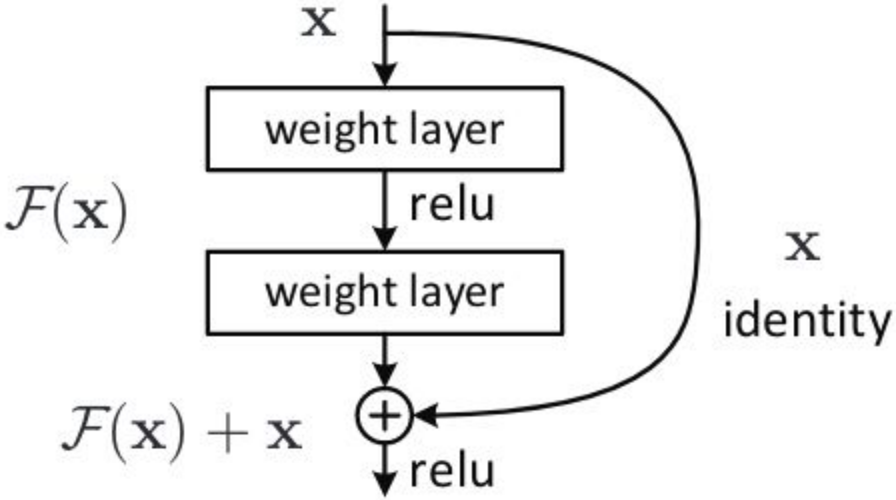
Automatic Relevance Determination

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33

Automatic Relevance Determination

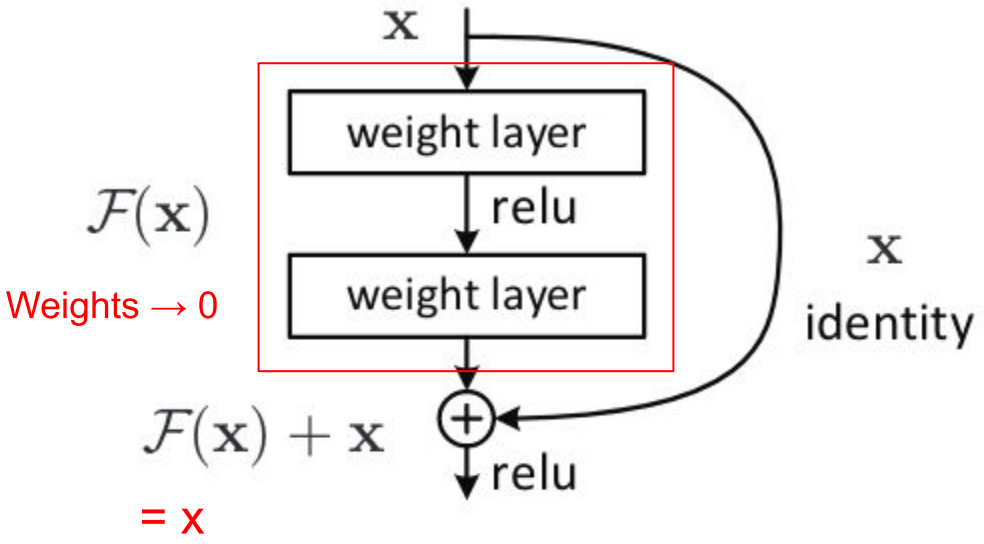
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
	-0.25	0.10	0.50	0.00	0.00	1.00	0.75	0.10	0.20	0.50	1.00	0.30	0.1
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33

ResNet



$\mathbf{x}$
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
$\mathbf{x}$
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

ResNet: Automatic Depth Determination



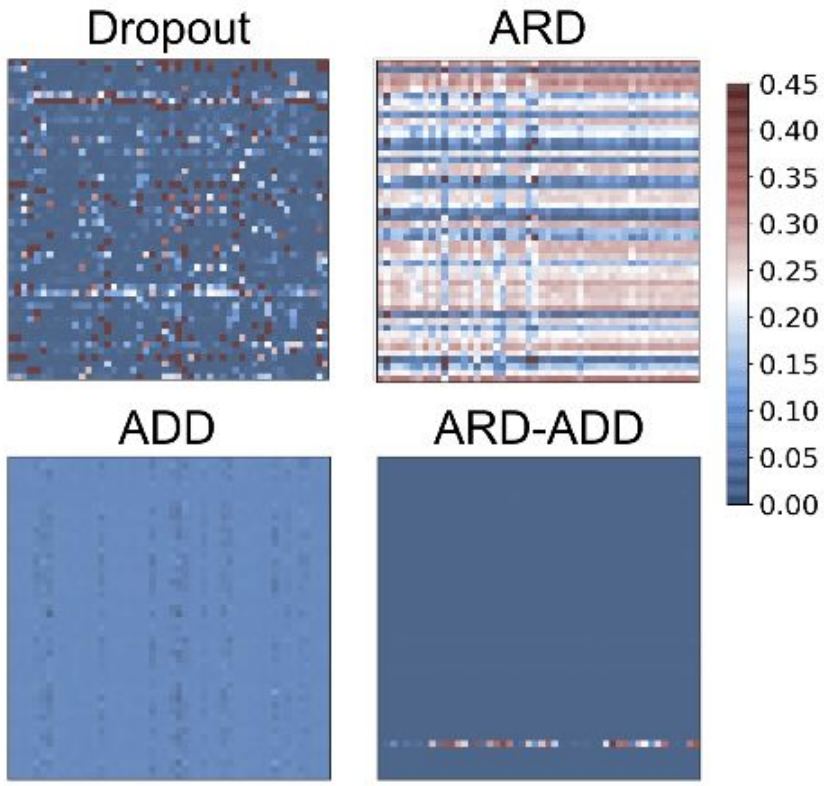


Figure 2. Posterior Structure.

Test Set RMSE

	Dropout	Prob. Backprop	Deep GP	ARD	ADD	ARD-ADD
Boston	$2.80 \pm .13$	$2.795 \pm .16$	$2.38 \pm .12$	$\mathbf{2.158 \pm .20}$	$2.343 \pm .31$	$2.367 \pm .18$
Concrete	$4.50 \pm .18$	$5.241 \pm .12$	$4.64 \pm .11$	$3.805 \pm .28$	$4.084 \pm .34$	$\mathbf{3.761 \pm .23}$
Energy	$\mathbf{0.47 \pm .01}$	$0.903 \pm .05$	$0.57 \pm .02$	$0.852 \pm .01$	$0.867 \pm .11$	$0.853 \pm .08$
Kin8nm	$0.08 \pm .00$	$0.071 \pm .00$	$\mathbf{0.05 \pm .00}$	$0.066 \pm .01$	$0.064 \pm .00$	$0.064 \pm .00$
Power	$3.63 \pm .04$	$4.028 \pm .03$	$3.60 \pm .03$	$3.486 \pm .10$	$3.290 \pm .06$	$\mathbf{3.236 \pm .07}$
Wine	$0.60 \pm .01$	$0.643 \pm .01$	$\mathbf{0.50 \pm .01}$	$0.561 \pm .03$	$0.555 \pm .01$	$0.538 \pm .03$
Yacht	$0.66 \pm .06$	$0.848 \pm .05$	$0.98 \pm .09$	$0.691 \pm .12$	$0.657 \pm .14$	$\mathbf{0.604 \pm .16}$
Avg. Rank	4.4 ± 1.7	5.6 ± 0.5	3.1 ± 1.8	3.0 ± 1.1	2.9 ± 10	$\mathbf{2.0 \pm 1.1}$