6.867 Machine Learning Fall 2017

Lecture 4. Bayesian Linear Regression

Two Envelope Game

- Rules
  - I choose two numbers and store them in envelope α and β
  - You get to choose an envelop, open it and read the number from it
  - Two options:
    - ONE. You can decide to stop there
    - TWO. You can choose to open the second envelop
  - Who Wins?
    - If the last number you read is the largest then you win else I win.

- *What is your chance of winning?*

An Elegant Solution

- Your algorithm
  - Randomly pick the first envelop and read it
  - Let the number from it be A
  - Randomly sample a number, say Q, as per $\mathcal{N}(0, 1)$
  - You decide to do
    - ONE. Stop there if Q < A
    - TWO. Choose to open the second envelop if Q > A

- Your chances of winning
  - More than 50%!

Moral of the Story

- It helps to view your question from *Bayesian* perspective
  - even if there is *nothing Bayesian* about it

- We'll follow this for linear regression today

  - Bayesian Linear Regression
  - Predictive Distribution
  - Equivalent Kernel Representation

From Lens of Model Selection

- Setup
  - Target ***Y***, Features / Attributes $\mathbf{X}$
- Decision theoretic view: minimize squared loss (or risk)

$$f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] \;\approx\; \mathbf{w}^T\mathbf{x}$$

- Maximum likelihood view: choose model that maximizes likelihood of data

$$Y = f(\mathbf{X}) + \varepsilon,\;\text{where}\;\varepsilon \;\sim\; \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$$

- Bayesian view: choose model whose likelihood is maximized

Bayesian Linear Regression

- Model likelihood

$$\mathbb{P}(\text{parameters}|\text{data}) \propto \mathbb{P}(\text{data}|\text{parameters}) \times \mathbb{P}(\text{parameters})$$

- Define prior of model parameters $\mathbf{w} \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$

$$\mathbb{P}(\text{parameter}) \propto \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \right)$$

- Data: $\mathbf{Y} = [y_n],\ \mathbf{X} = [x_{ni}, 0 \leq i \leq p],\ 1 \leq n \leq N$

$$\mathbb{P}(\text{data}|\text{parameter}) \propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\mathbf{w})^T (\mathbf{Y} - \mathbf{X}\mathbf{w}) \right)$$

Bayesian Linear Regression

- Posterior on model parameters

$$
\begin{aligned}
&\mathbb{P}(\text{parameter}|\text{data}) \\
&\propto \exp\Big(-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\mathbf{w})^T(\mathbf{Y}-\mathbf{X}\mathbf{w}) - \frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\mathbf{w}^T(\mathbf{S}_0^{-1}+\sigma^{-2}\mathbf{X}^T\mathbf{X})\mathbf{w} + (\mathbf{S}_0^{-1}\mathbf{m}_0 + \sigma^{-2}\mathbf{Y}^T X)\mathbf{w}\Big) \\
&\propto \exp\Big(-\frac{1}{2}\mathbf{w}^T\mathbf{J}\mathbf{w} + \mathbf{h}^T\mathbf{w}\Big)
\end{aligned}
$$

- where

$$
\begin{aligned}
\mathbf{J} &= (\mathbf{S}_0^{-1}+\sigma^{-2}\mathbf{X}^T\mathbf{X}) \\
\mathbf{h} &= (\mathbf{S}_0^{-1}\mathbf{m}_0 + \sigma^{-2}\mathbf{Y}^T X)
\end{aligned}
$$

Bayesian Linear Regression

- Gaussian Distribution: Equivalent Forms

Standard Form Information Form

$$\mathcal{N}(\mu, \Sigma) \;\Leftrightarrow \mathcal{N}^{-1}(h, J)$$

$$\exp\Big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big) \;\Leftrightarrow\; \exp\Big(-\frac{1}{2}x^TJx + h^Tx\Big)$$

$$\Sigma^{-1} = J$$

$$\mu = J^{-1}h$$

Bayesian Linear Regression

- Posterior on model parameters

$$\mathbf{J} = (\mathbf{S}_0^{-1} + \sigma^{-2}\mathbf{X}^T\mathbf{X})$$
$$\mathbf{h} = (\mathbf{S}_0^{-1}\mathbf{m}_0 + \sigma^{-2}\mathbf{Y}^T X)$$

- That is, model parameter has Gaussian distribution with parameters

$$\mathbf{S}_N^{-1} = (\mathbf{S}_0^{-1} + \sigma^{-2}\mathbf{X}^T\mathbf{X})$$
$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sigma^{-2}\mathbf{Y}^T X)$$

- Since its Gaussian, the mode of distribution over parameters is the mean
  - Answer to Bayesian Linear Regression

Bayesian Linear Regression

- An example: $\mathbf{S}_0 = \beta^2 \mathbf{I}, \;\; \mathbf{m}_0 = \mathbf{0}$

$$\mathbf{S}_N^{-1} = (\beta^{-2}\mathbf{I} + \sigma^{-2}\mathbf{X}^T\mathbf{X})$$

$$\mathbf{m}_N = \left(\frac{\sigma^2}{\beta^2}\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{Y}^T X$$

- The corresponding log-posterior

$$\log \mathbb{P}(\mathbf{w}|\mathrm{data}) = -\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\mathbf{w})^T(\mathbf{Y} - \mathbf{X}\mathbf{w}) - \frac{1}{2\beta^2}\mathbf{w}^T\mathbf{w} + \mathrm{const.}$$

- Maximizing this posterior is same as solving *Ridge Regression!*
  - That is, we've found Bayesian view of Ridge Regression

An Experiment

- Data generation:

$$x_n \sim U[-1, 1]$$

$$y_n = a_0 + a_1 x_n + \varepsilon_n$$

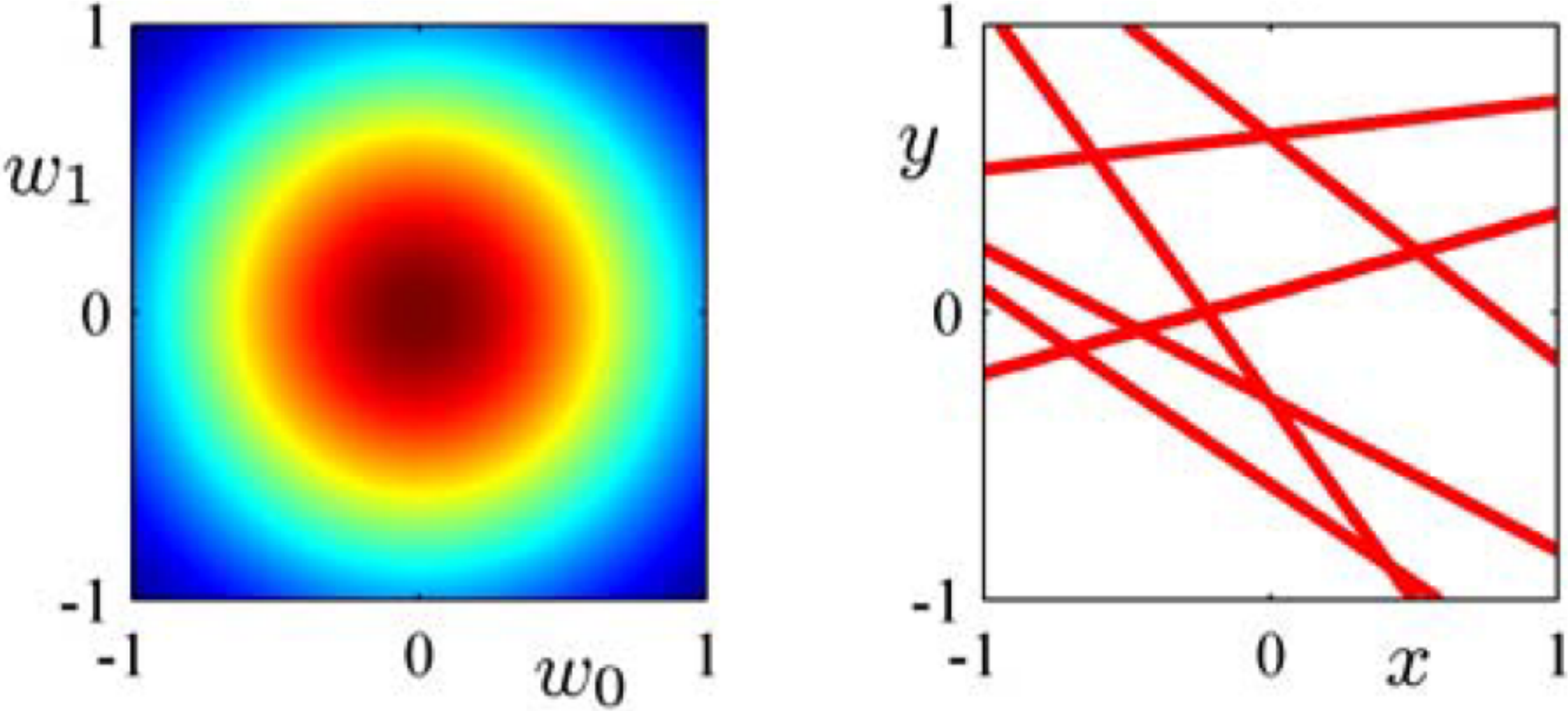
$$\text{where } a_0 = -0.3,\ a_1 = 0.5,\ \varepsilon_n \sim \mathcal{N}(0, 0.2^2)$$

- We want to learn parameters $\mathbf{w} = [w_0, w_1]$

- We utilize prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, 0.5\mathbf{I})$

An Experiment

- Posterior and Model after

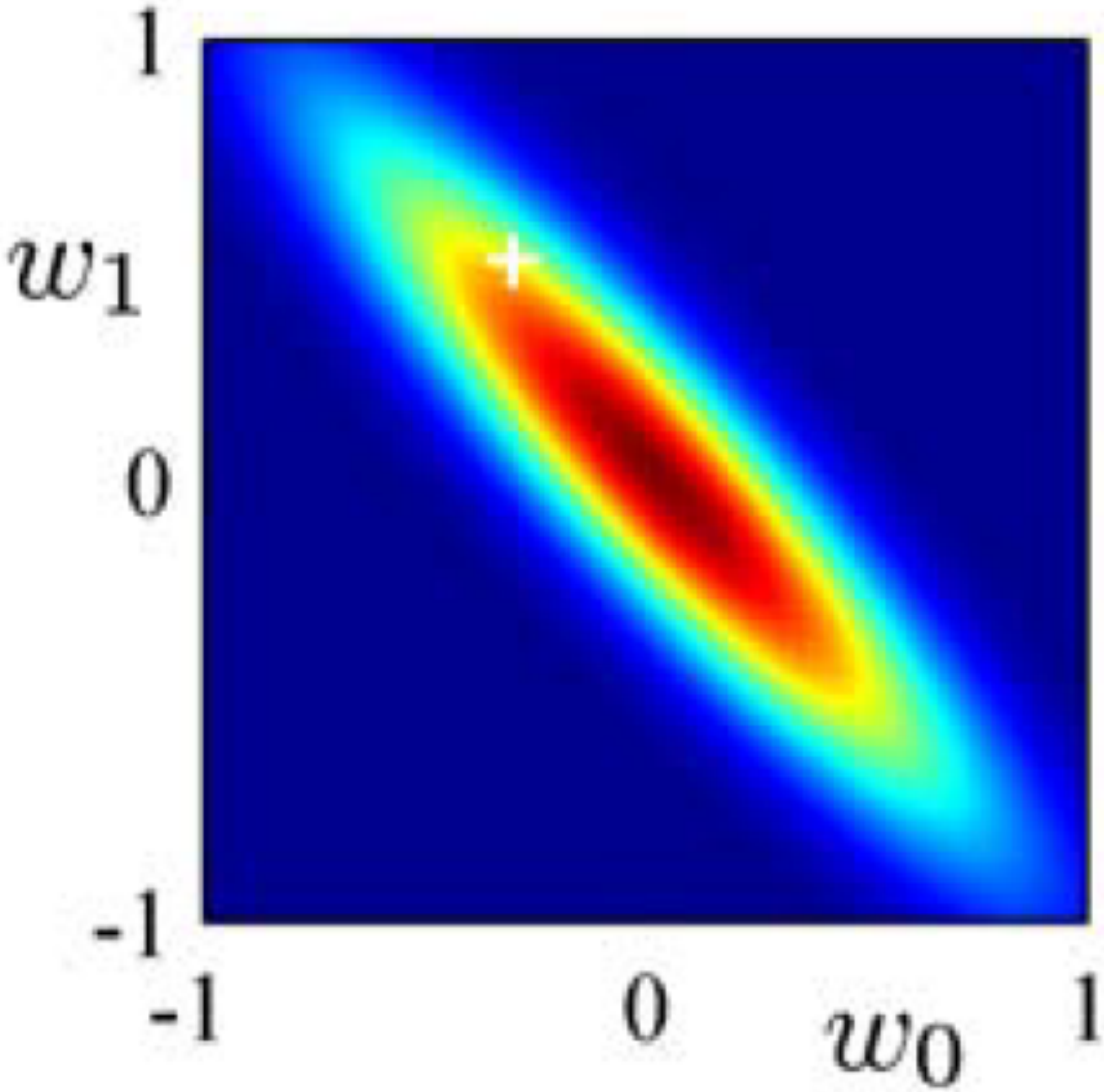


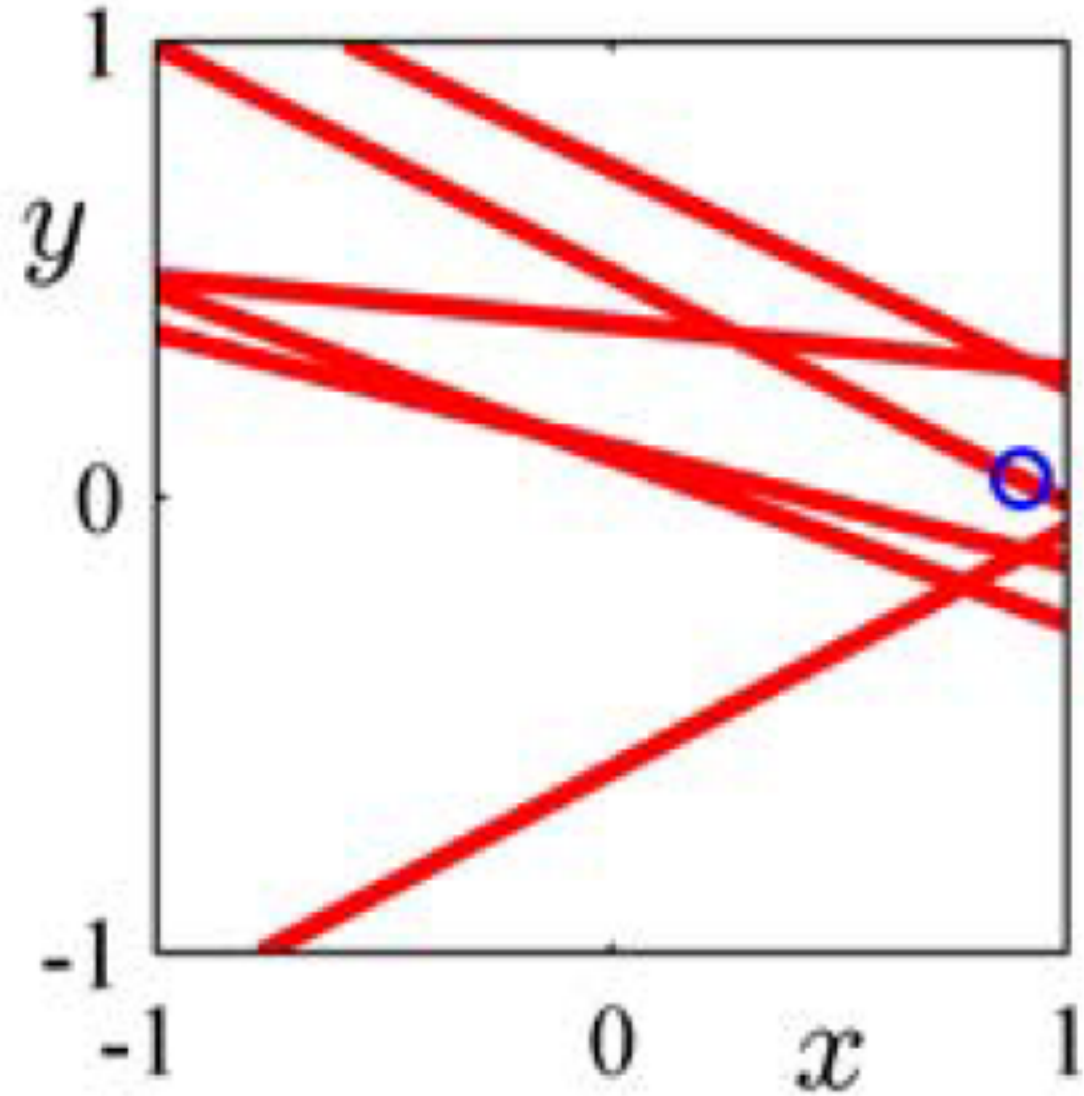
Samples = 0

Images: Courtsey C. Bishop, *Pattern Recognition and Machine Learning*

An Experiment

- Posterior and Model after



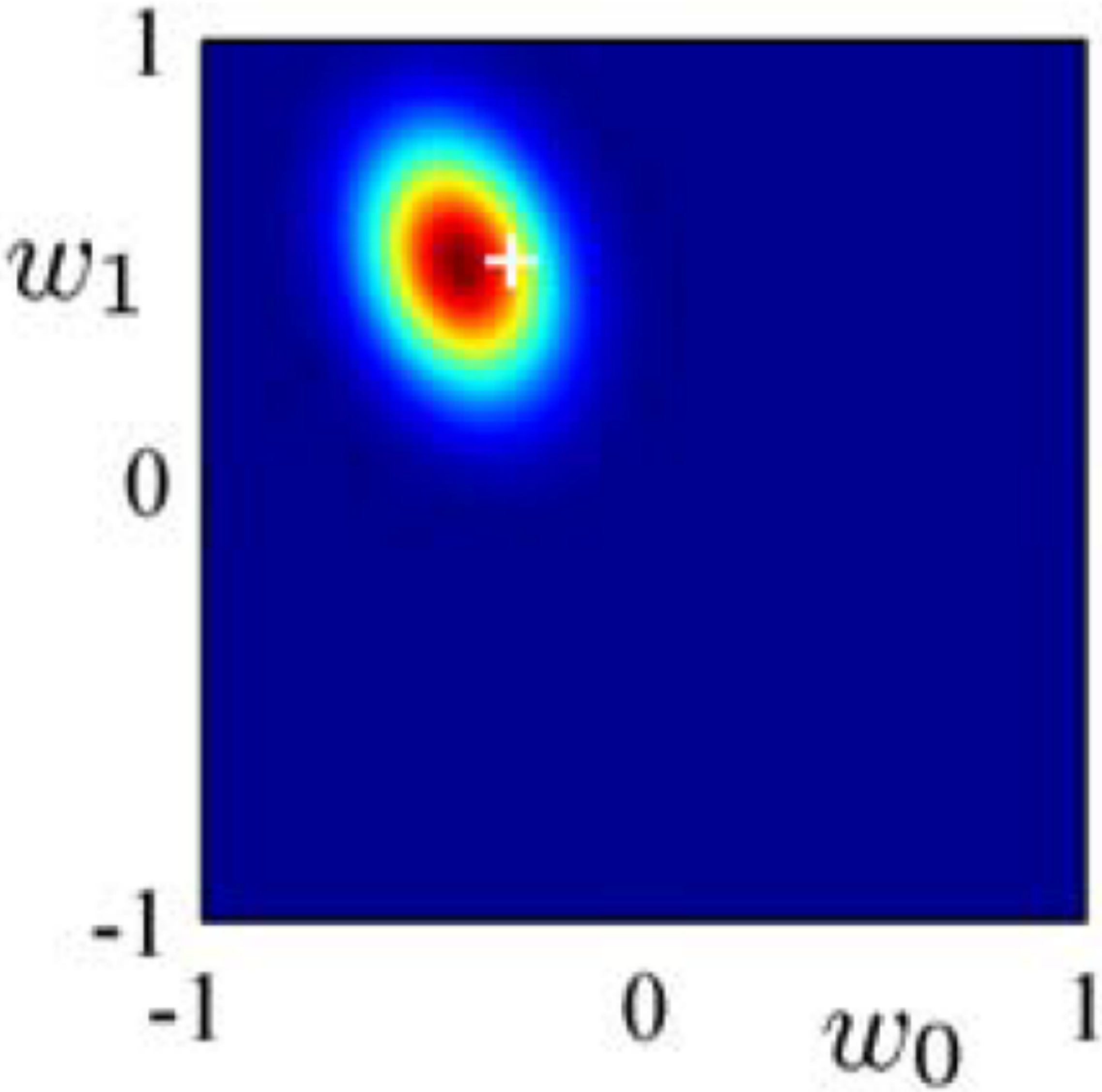


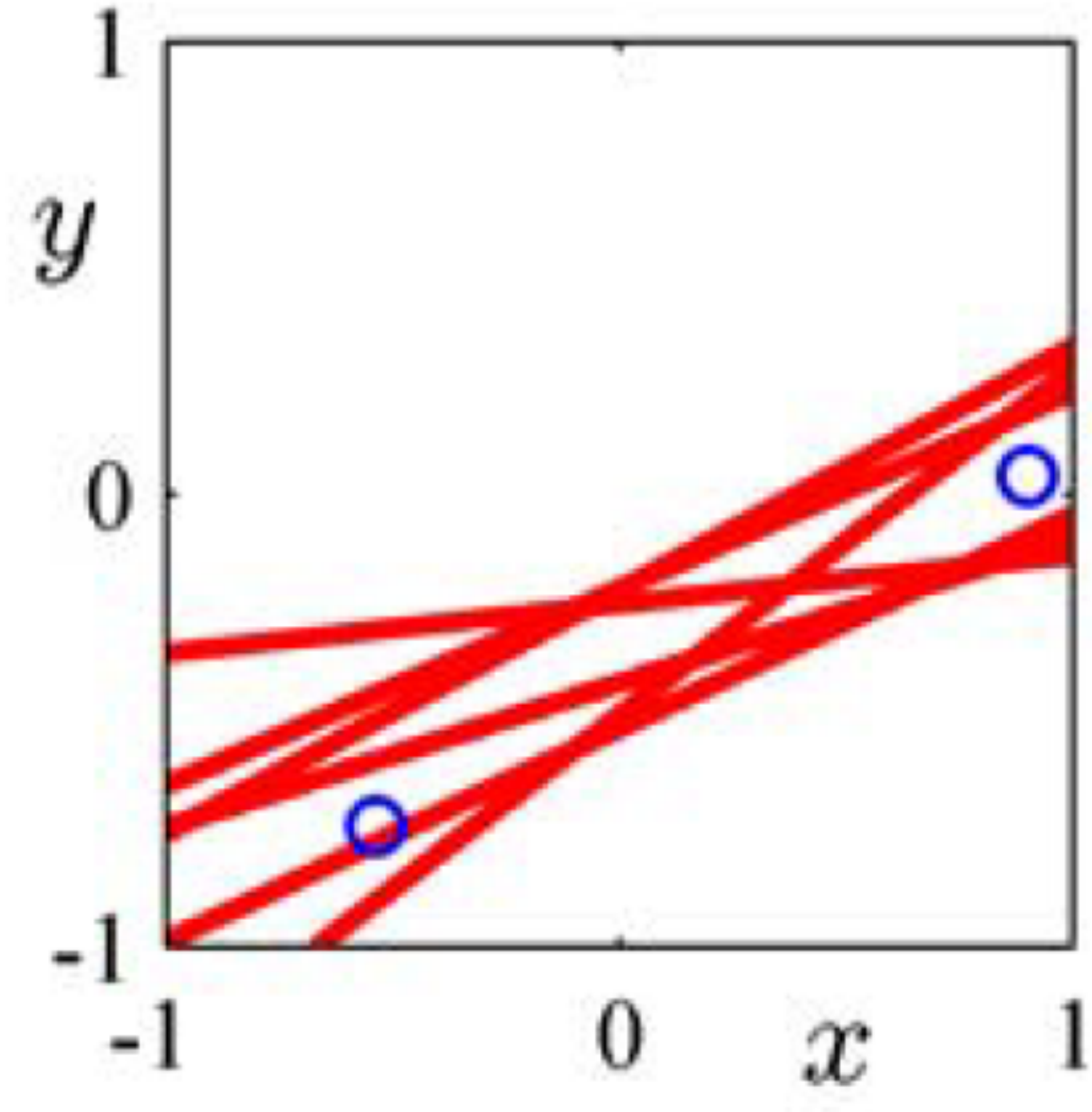
Samples = 1

Images: Courtsey C. Bishop, *Pattern Recognition and Machine Learning*

An Experiment

- Posterior and Model after



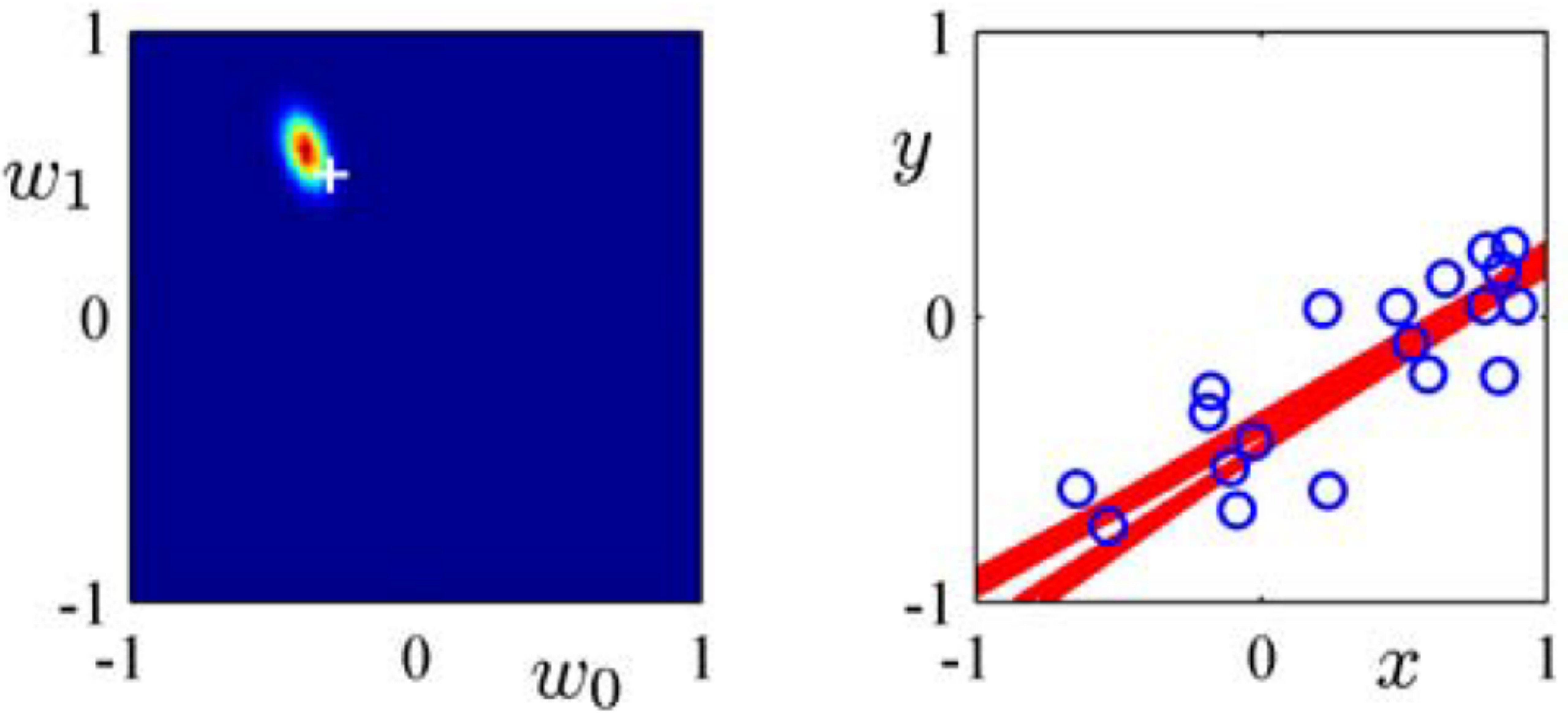


Samples = 2

Images: Courtsey C. Bishop, *Pattern Recognition and Machine Learning*

An Experiment

- Posterior and Model after



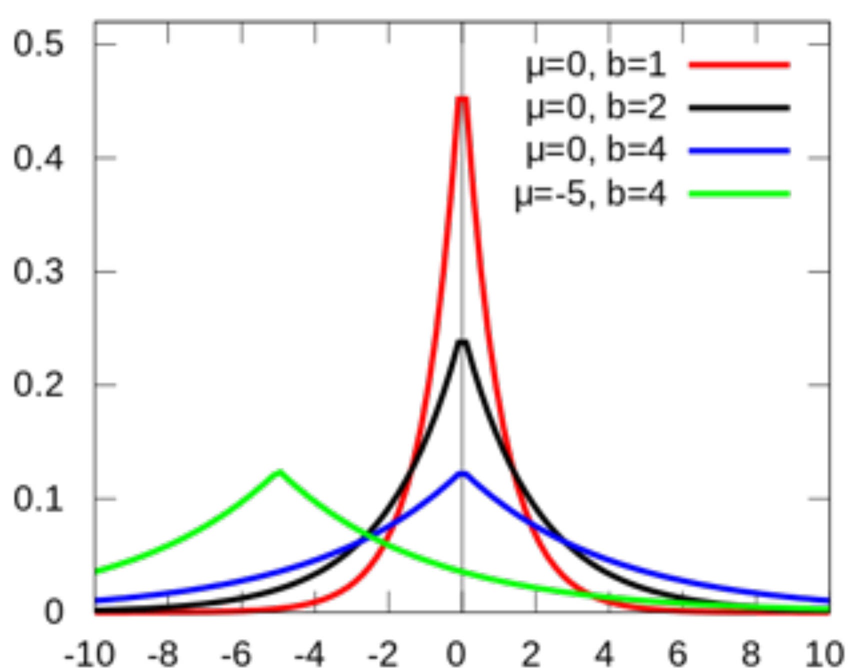
Samples = 20

Images: Courtsey C. Bishop, *Pattern Recognition and Machine Learning*

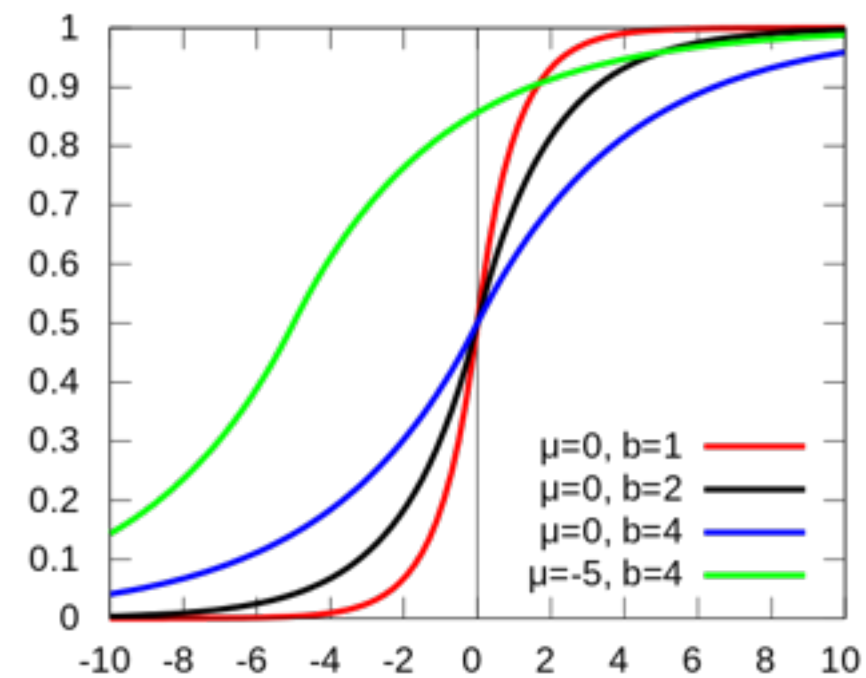
Bayesian Linear Regression

- Gaussian prior leads to Ridge regression
- Other priors lead to different regularization, e.g.
  - LASSO corresponds to Laplace Prior

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{2b} \right)$$



PDF



CDF

Predictive Distribution

- Bayesian view immediately provides distribution over
  - Model parameter
  - Target variable for given Feature / Attribute: called *predictive distribution*

$$\mathbb{P}(y|x, \text{data}) = \int \mathbb{P}(y|\mathbf{w}, x)\mathbb{P}(\mathbf{w}|\text{data}) d\mathbf{w}$$

  - Now

$$y|\mathbf{w} \sim \mathcal{N}(\mathbf{w}^T x, \sigma^2) \quad \text{and} \quad \mathbf{w} \sim \mathcal{N}(\mathbf{m}_N, S_N)$$

  - That is, y is Gaussian

Predictive Distribution

- Mean of y

$$\mathbb{E}[y] = \mathbb{E}[\mathbb{E}[y|\mathbf{w}]] \ = \ \mathbb{E}[\mathbf{w}^T x]$$
$$= \mathbb{E}[\mathbf{w}]^T x \ = \ \mathbf{m}_N^T x$$

- Now

$$\begin{aligned}
\mathbb{E}[y^2] &= \mathbb{E}[\mathbb{E}[y^2|\mathbf{w}]] \\
&= \mathbb{E}[\mathbb{E}[(\mathbf{w}^T x + \varepsilon)^2|\mathbf{w}]] \text{ where } \varepsilon \sim \mathcal{N}(0, \sigma^2) \\
&= \mathbb{E}[\mathbb{E}[x^T \mathbf{w}\mathbf{w}^T x|\mathbf{w}]] + \sigma^2 \\
&= x^T(\mathbf{m}_N \mathbf{m}_N^T + \mathbf{S}_N)x + \sigma^2, \ \text{since} \mathbf{w} \sim \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N)
\end{aligned}$$

- Therefore, variance of y

$$\sigma_N^2(x) \equiv \mathrm{Var}[y] = x^T \mathbf{S}_N x + \sigma^2$$

Equivalent Kernel

- Re-writing mean of predictive y (for given attribute x)

$$y(x, \mathbf{m}_N) = \mathbf{m}_N^T x$$

- Recall with $\mathbf{m}_0 = \mathbf{0}$

$$\mathbf{m}_N = \sigma^{-2}\mathbf{S}_N\mathbf{X}^T\mathbf{Y}$$

- Therefore

$$\begin{aligned} y(x, \mathbf{m}_N) &= \sigma^{-2}x^T\mathbf{S}_N\mathbf{X}^T\mathbf{Y} \\ &= \sigma^{-2}\Big(\sum_n x^T\mathbf{S}_N x_n y_n\Big) \\ &= \sum_n k_N(x, x_n) y_n \end{aligned}$$

- Where $k_N(x, x_n) = \sigma^{-2}x^T\mathbf{S}_N x_n$

Equivalent Kernel

- Observe that

$$\begin{aligned}\mathrm{Cov}[y(x), y(x')] &= \mathrm{Cov}[w^T x, w^T x'] \\ &= x^T \mathrm{Cov}[w, w] x' \\ &= x^T \mathbf{S}_N x' \\ &= \sigma^2 k_N(x, x')\end{aligned}$$

- That is

$$k_N(x, x') = \psi(x)^T \psi(x')$$

- Where

$$\psi(x) = \sigma^{-1} \mathbf{S}_N^{1/2} x$$