# Computazione per l'interazione naturale: Regressione lineare Bayesiana

Corso di Interazione uomo-macchina II

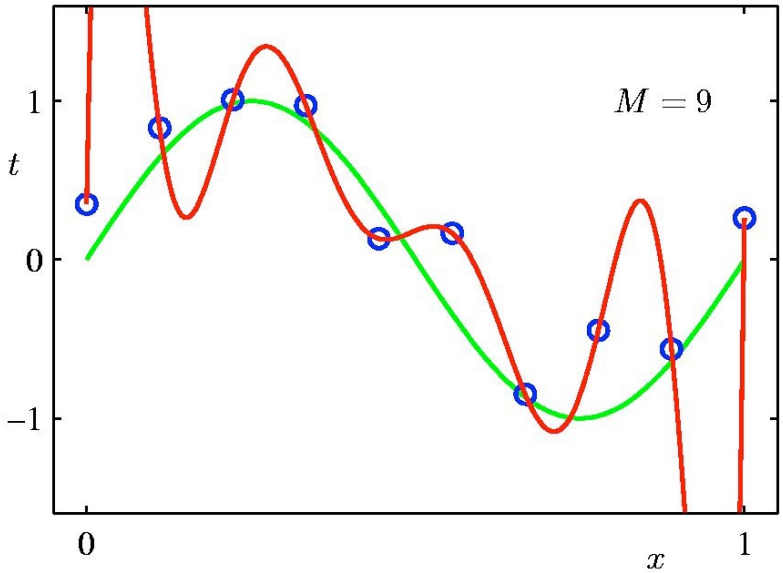
Prof. Giuseppe Boccignone

Dipartimento di Informatica
Università di Milano

boccignone@di.unimi.it
http://boccignone.di.unimi.it/IUM2_2014.html

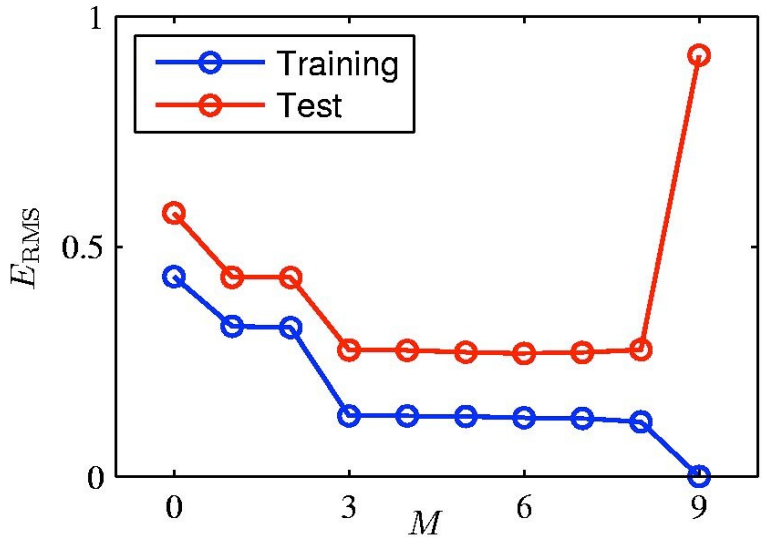
## Regressione lineare (deterministica) //Il problema della regolarizzazione (deterministica)

overfitting!




Polinomio di ordine 9

# Regressione lineare (deterministica)
# //Il problema della regolarizzazione (deterministica)




Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

# Regressione lineare (deterministica)
# //Il problema della regolarizzazione (deterministica)

| | $M=0$ | $M=1$ | $M=3$ | $M=9$ |
|---|---|---|---|---|
| $w_0^\star$ | 0.19 | 0.82 | 0.31 | 0.35 |
| $w_1^\star$ | | -1.27 | 7.99 | 232.37 |
| $w_2^\star$ | | | -25.43 | -5321.83 |
| $w_3^\star$ | | | 17.37 | 48568.31 |
| $w_4^\star$ | | | | -231639.30 |
| $w_5^\star$ | | | | 640042.26 |
| $w_6^\star$ | | | | -1061800.52 |
| $w_7^\star$ | | | | 1042400.18 |
| $w_8^\star$ | | | | -557682.99 |
| $w_9^\star$ | | | | 125201.43 |

# Regressione lineare (deterministica) //Il problema della regolarizzazione (deterministica)

- Introduzione di un termine di regolarizzazione nella funzione di costo

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

dipendendente dai dati

dipendendente dai parametri

regularization parameter

- Esempio: funzione quadratica che penalizza grandi coefficienti

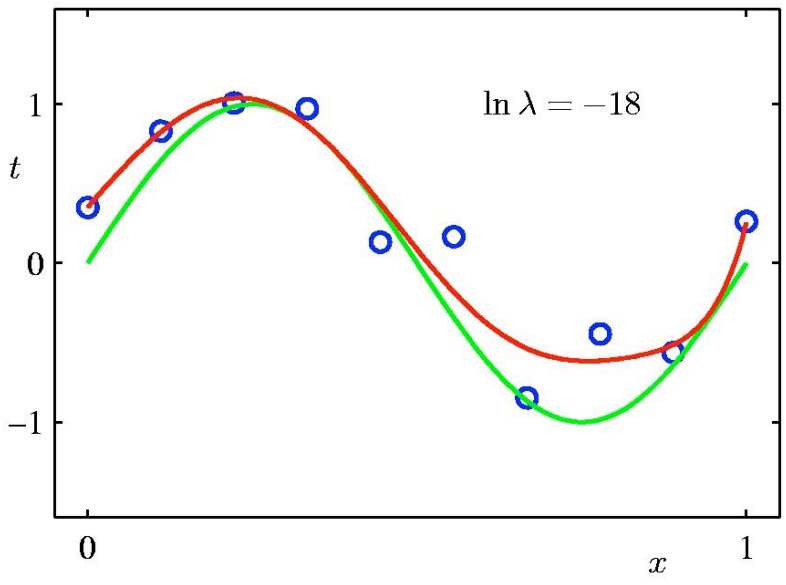
$$E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{2}\sum_{i=0}^{M-1} w_i^2$$

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i-1}^{N}\{t_i - \mathbf{w}^T\phi(\mathbf{x}_i)\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

penalized loss function

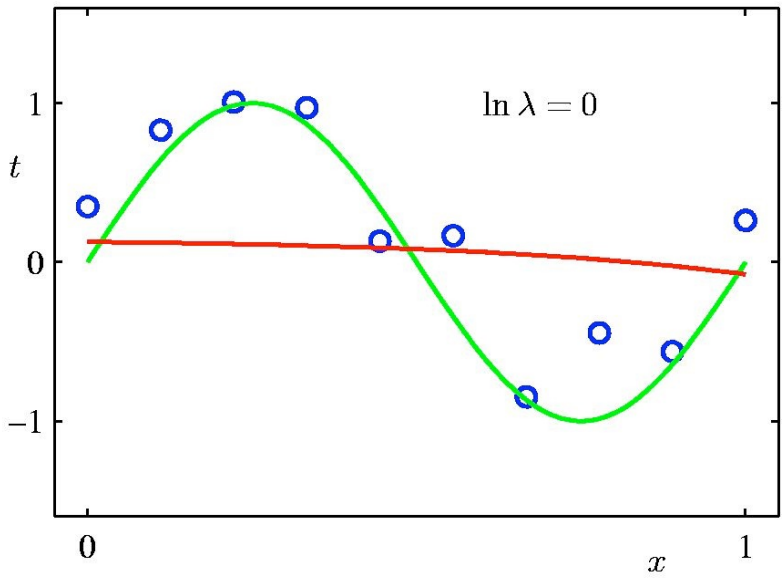
$$\mathbf{w} = (\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$$

# Regressione lineare (deterministica) //Il problema della regolarizzazione (deterministica)



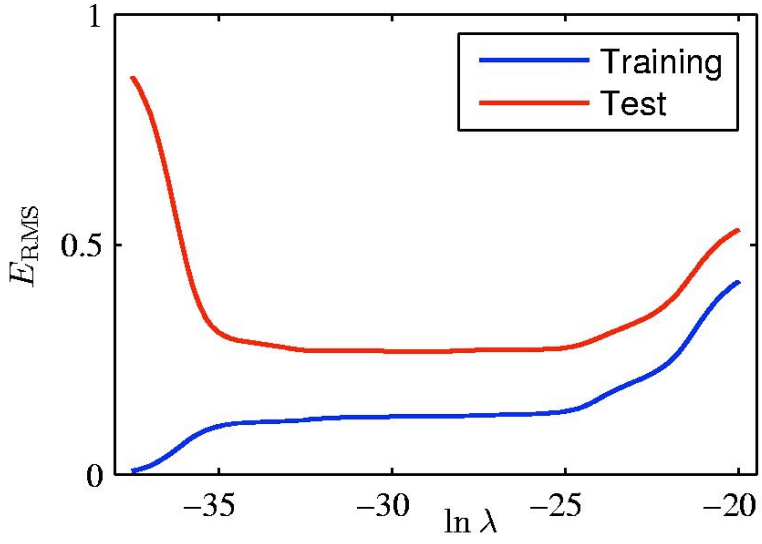

$$\ln \lambda = -18$$

## Regressione lineare (deterministica)
## //Il problema della regolarizzazione (deterministica)




$\ln \lambda = 0$

## Regressione lineare (deterministica)
## //Il problema della regolarizzazione (deterministica)




$E_{\mathrm{RMS}}$ vs $\ln \lambda$

# Regressione lineare (deterministica) //Il problema della regolarizzazione (deterministica)

| | $\ln \lambda = -\infty$ | $\ln \lambda = -18$ | $\ln \lambda = 0$ |
|---|---|---|---|
| $w_0^\star$ | 0.35 | 0.35 | 0.13 |
| $w_1^\star$ | 232.37 | 4.74 | -0.05 |
| $w_2^\star$ | -5321.83 | -0.77 | -0.06 |
| $w_3^\star$ | 48568.31 | -31.97 | -0.05 |
| $w_4^\star$ | -231639.30 | -3.89 | -0.03 |
| $w_5^\star$ | 640042.26 | 55.28 | -0.02 |
| $w_6^\star$ | -1061800.52 | 41.32 | -0.01 |
| $w_7^\star$ | 1042400.18 | -45.95 | -0.00 |
| $w_8^\star$ | -557682.99 | -91.53 | 0.00 |
| $w_9^\star$ | 125201.43 | 72.68 | 0.01 |

# Regressione lineare probabilistica ML //problemi

- Valore pratico limitato: ci vogliono molti data sets (ensemble di data set)
- Approccio frequentistico
- In generale:
  - L'utilizzo del criterio di massima verosimiglianza (ML) per la determinazione dei parametri del modello tende tipicamente ad essere soggetto al problema dell'overfitting
  - Per controllare la complessità del modello, un approccio di tipo bayesiano, invece di fare uso del termine di regolazione E(w), introduce una distribuzione che rappresenta la nostra valutazione a priori (prima dell'osservazione del training set) dei valori dei coefficienti w

# Esempio: stima di massima verosimiglianza // lancio di moneta

- Per un singolo lancio uso Bernoulli

$$p(l|\theta) = \theta^{I(l=T)}(1-\theta)^{I(l=C)} \qquad I(l=T)$$

1 se il lancio ha dato "testa" come esito
0 altrimenti

- Likelihood per N lanci indipendenti

$$p(S|\theta) = \prod_{i=1}^{N} \theta^{I(l=T)}(1-\theta)^{I(l=C)} = \theta^{T}(1-\theta)^{N-T}$$

- Stima ML

$$\frac{dp(S|\theta)}{d\theta} = T\theta^{T-1}(1-\theta)^{N-T} - (N-T)\theta^{T}(1-\theta)^{N-T-1} = 0$$

$$\Rightarrow \quad \boxed{\hat{\theta} = \frac{T}{N}}$$

- Quindi se dopo N=3 lanci osservo S = {C, C, C} allora

$$\hat{\theta} = \frac{T}{N} = 0 \qquad \boxed{:\text{-(}}$$

# Esempio: stima Bayesiana // lancio di moneta

- Nel setting ML (frequentistico) in realtà con N piccolo ottengo conclusioni fuorvianti (la moneta è sicuramente truccata: darà sempre CCCCCCC....)

- Approccio Bayesiano:

a posteriori: distribuzione Beta

$$f(\theta|S) = \frac{p(S|\theta)f(\theta)}{p(S)}$$

funzione di likelihood: distribuzione di Bernoulli

a priori coniugata di Bernoulli: distribuzione Beta

# Esempio: stima Bayesiana // lancio di moneta

- Approccio Bayesiano:

funzione di likelihood: distribuzione di Bernoulli

a priori coniugata di Bernoulli: distribuzione Beta

a posteriori: distribuzione Beta

$$f(\theta|S) = \frac{p(S|\theta)f(\theta)}{p(S)}$$

a priori coniugata di Bernoulli: distribuzione Beta

$$f(\theta|\alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} \propto \theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$$

funzione di likelihood: distribuzione di Bernoulli

$$p(S|\theta) = \prod_{i=1}^{N}\theta^{I(l=T)}(1-\theta)^{I(l=C)} = \theta^{T}(1-\theta)^{N-T}$$

a posteriori: distribuzione Beta

$$f(\theta|S) \propto \theta^{T+\alpha_1-1}(1-\theta)^{N-T+\alpha_2-1} \propto Beta(T+\alpha_1, N-T+\alpha_2)$$

# Esempio: stima Bayesiana // lancio di moneta

- Approccio Bayesiano:

funzione di likelihood: distribuzione di Bernoulli

a priori coniugata di Bernoulli: distribuzione Beta

a posteriori: distribuzione Beta

$$f(\theta|S) = \frac{p(S|\theta)f(\theta)}{p(S)}$$

a posteriori: distribuzione Beta

$$f(\theta|S) \propto \theta^{T+\alpha_1-1}(1-\theta)^{N-T+\alpha_2-1} \propto Beta(T+\alpha_1, N-T+\alpha_2)$$

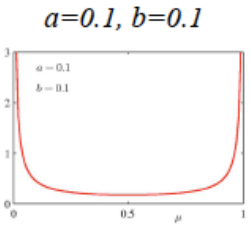
- Uso come stima il valore atteso della Beta

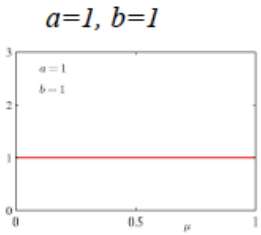
$$E[\theta|S] = \frac{T+\alpha_1}{N+\alpha_1+\alpha_2}$$

# Esempio: stima Bayesiana // lancio di moneta

- Uso come stima il valore atteso della Beta

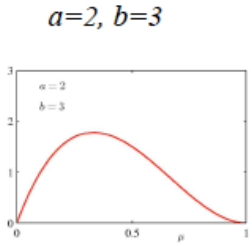
$$E[\theta|S] = \frac{T + \alpha_1}{N + \alpha_1 + \alpha_2}$$

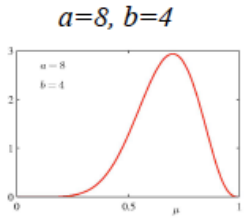



- Anche se uso un a priori uniforme $\alpha_1 = \alpha_2 = 1$

$$\hat{\theta} = \frac{T + 1}{N + 2}$$




- Quindi se dopo N=3 lanci osservo S = {C, C, C} allora

$$\hat{\theta} = \frac{T + 1}{N + 2} = 0.2$$

;-)

# Esempio: stima Bayesiana // lancio di moneta

- Non a caso abbiamo utilizzato come stima il valore atteso della Beta

$$E[\theta|S] = \frac{T + \alpha_1}{N + \alpha_1 + \alpha_2}$$

- Implicitamente abbiamo usato lo Step 3: **predizione Bayesiana**:

$$p(x|S) = \int p(x|\theta)p(\theta|S)d\theta$$

- Nel nostro esempio

$$p(X = 1|S) = \int_0^1 p(X = 1|\theta)p(\theta|S)d\theta = \int_0^1 \theta \text{Beta}(\theta|\alpha_1', \alpha_2')d\theta = E[\theta] = \frac{\alpha_1'}{\alpha_2'}$$

$$\alpha_1' = \alpha_1 + T \text{ e } \alpha_2' = \alpha_1 + \alpha_2 + N$$

# Esempio: stima Bayesiana // lancio di moneta

- Qual è la relazione con la stima di massima verosimiglianza?
- Per infiniti lanci

$$p(\theta|S) \rightarrow \delta(\theta - \hat{\theta})$$

- ovvero:

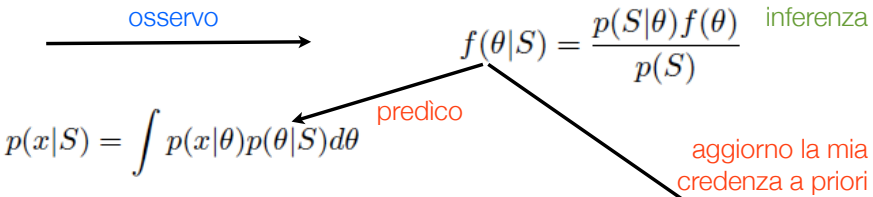
$$p(X=1|S) = \int_0^1 p(X=1|\theta)p(\theta|S)d\theta \approx \int_0^1 p(X=1|\theta)\delta(\theta-\hat{\theta})d\theta = p(X=1|\hat{\theta})$$

- Infatti per T e N >> $\alpha_1$ e $\alpha_2$

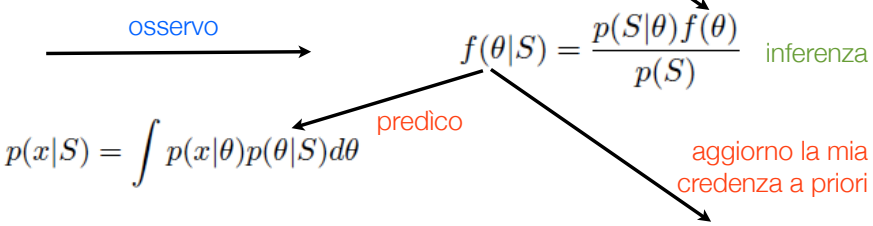
$$E[\theta|S] = \frac{T+\alpha_1}{N+\alpha_1+\alpha_2} \longrightarrow \hat{\theta} = \frac{T}{N}$$

# Esempio: stima Bayesiana // lancio di moneta

- Posso usare la formula di Bayes in modo ricorsivo (sequenziale)
  - ad ogni lancio, osservo T o C e uso come a priori la posteriori (Beta) del lancio precedente, quindi effettuo la predizione
    - S={C,C,C}




    - S={C,C,C,T}



# Esempio: stima Bayesiana
# // lancio di moneta

- Nel nostro modello h possiamo calcolare in forma chiusa anche la verosimiglianza marginale o evidenza (dei dati)

$$p(\theta|S,h) = \frac{p(S|\theta,h)p(\theta|h)}{p(S|h)} \propto p(S|\theta,h)p(\theta|h)$$

- Ricordando che

$$p(\theta|S) = \text{Beta}(\alpha_1',\alpha_2') = \frac{\Gamma(\alpha_1'+\alpha_2')}{\Gamma(\alpha_1')\Gamma(\alpha_2')}\theta^{\alpha_1'-1}(1-\theta)^{\alpha_2'-1}$$

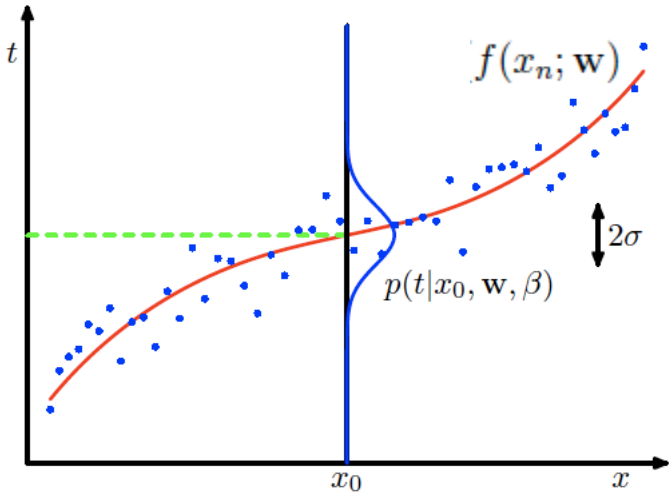
- Usando Bayes

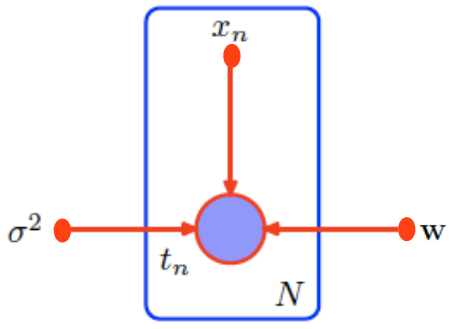
$$\begin{aligned} p(\theta|S) &= \frac{p(\theta)p(S|\theta)}{p(S)} \\ &= \frac{1}{p(S)}\left[\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}\right][\theta^T(1-\theta)^{N-T}] \\ &= \frac{1}{p(S)}\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta^{\alpha_1'-1}(1-\theta)^{\alpha_2'-1} \end{aligned}$$

$$\Rightarrow \quad \boxed{p(S) = \frac{\Gamma(\alpha_1')\Gamma(\alpha_2')}{\Gamma(\alpha_1'+\alpha_2')}\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}}$$

# Regressione lineare
# //Approccio Bayesiano: sintesi

- Framework probabilistico $p(\mathbf{t}|\mathbf{x},\mathbf{w},\sigma) = \prod_{n=1}^{N} p(t_n|x_n) = \prod_{n=1}^{N} \mathcal{N}(f(x_n;\mathbf{w}),\sigma)$



# Regressione lineare //Approccio Bayesiano: sintesi

- Framework probabilistico $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma) = \prod_{n=1}^{N} p(t_n|x_n) = \prod_{n=1}^{N} \mathcal{N}(f(x_n; \mathbf{w}), \sigma)$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

Regola di Bayes

$$p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \sigma) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)p(\mathbf{w}) = p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma)p(\mathbf{t}|\mathbf{X}, \sigma)$$

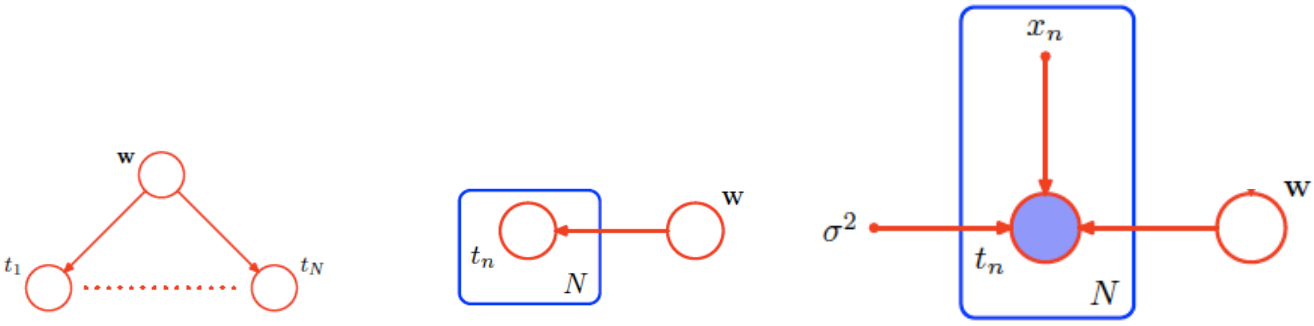
Probabilità congiunta

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)\frac{p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X}, \sigma)}$$

Regola di Bayes per la regressione

# Regressione lineare //Approccio Bayesiano: sintesi

- I pesi **w** non sono semplici parametri ma variabili aleatorie su cui fare inferenza abbiamo cioè un modello generativo

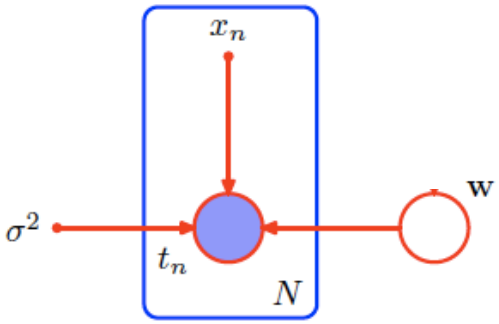



$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$

# Regressione lineare //Approccio Bayesiano: sintesi

- I pesi **w** non sono semplici parametri ma variabili aleatorie su cui fare inferenza abbiamo cioè un modello generativo

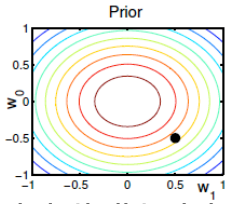
$$p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \sigma) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma)p(\mathbf{w})$$




Probabilità congiunta
(probabilità di tutto)

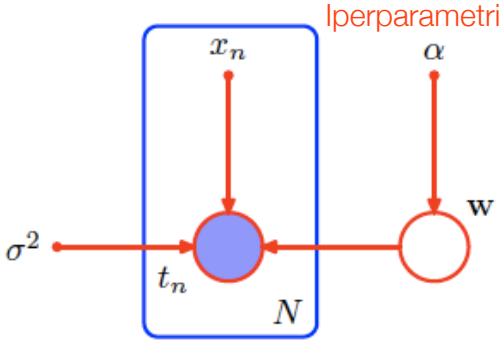
# Regressione lineare //Approccio Bayesiano: sintesi

- I pesi **w** non sono semplici parametri ma variabili aleatorie su cui fare inferenza abbiamo cioè un modello generativo
  - Si introduce una probabilità a priori su **w** (che regolarizza: favorisce piccoli **w**)




$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

Iperparametri



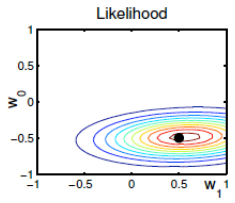

- Dai dati di training si calcola la probabilità a posteriori

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) \times p(\mathbf{w}|\alpha)$$

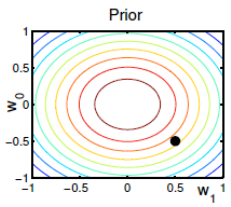
$$\mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}) \times \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

# Regressione lineare //Approccio Bayesiano: sintesi

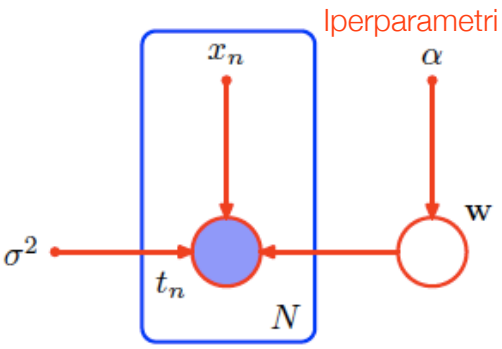
- Dai dati di training si calcola la probabilità a posteriori

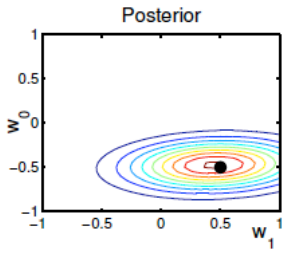



$\mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1})$ $\times$




$\mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I})$




$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\alpha,\beta) \propto p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) \times p(\mathbf{w}|\alpha)$$

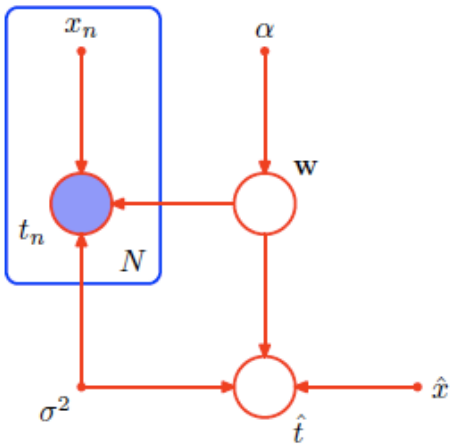
# Regressione lineare //Approccio Bayesiano: sintesi

- Dai dati di training si calcola la probabilità a posteriori

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\alpha,\beta) \propto p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) \times p(\mathbf{w}|\alpha)$$

- Per un nuovo dato **x**, predico t usando la distribuzione predittiva

$$p(t|\mathbf{x},\mathbf{X},\mathbf{t},\alpha,\beta) = \int p(t|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\mathbf{X},\alpha,\beta)d\mathbf{w}$$

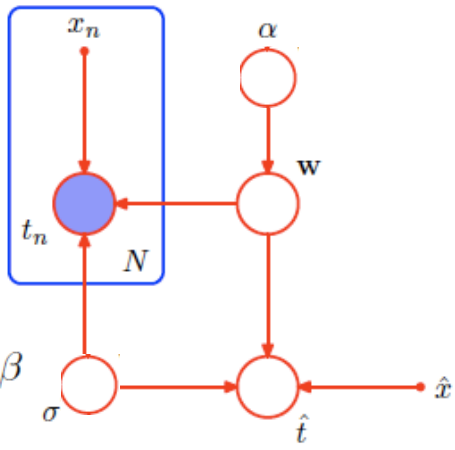



$$\hat{t} \perp t_n|\mathbf{w}$$

# Regressione lineare //Approccio Bayesiano: sintesi

- In un approccio totalmente Bayesiano anche gli iper-parametri sono variabili aleatorie

Variabili aleatorie




- Per un nuovo dato **x**, predico t usando la distribuzione predittiva

$$p(t|\mathbf{x},\mathbf{X},\mathbf{t}) =$$

$$\int p(t|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\mathbf{X},\alpha,\beta)p(\alpha,\beta|\mathbf{X},\mathbf{t})d\mathbf{w}d\alpha d\beta$$

- problema intrattabile in forma chiusa!!

$$\hat{t} \perp t_n|\mathbf{w}$$

Modello Bayesiano gerarchico

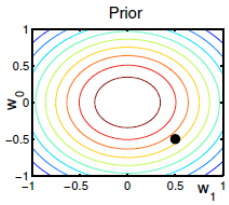
# Regressione lineare //Approccio Bayesiano: l'ipotesi sul prior

- Si assume che i coefficienti w siano tra loro indipendenti e tutti distribuiti allo stesso modo, secondo una gaussiana a media 0 e varianza $\alpha^{-1}$

$$p(\mathbf{w}|\alpha) = \prod_{i=0}^{M-1} \left(\frac{\alpha}{2\pi}\right)^{1/2} exp\left(-\frac{\alpha}{2}w_i^2\right)$$

- ovvero una gaussiana multivariata

$$f(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$




- con parametri

$$\mu = 0 \qquad \Sigma = \begin{bmatrix} \alpha^{-1} & 0 & \cdots & 0 \\ 0 & \alpha^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha^{-1} \end{bmatrix}$$

# Regressione lineare
# //Approccio Bayesiano: l'ipotesi sul prior

- Sappiamo che (the Matrix cookbook)

$$\mathbf{x} \sim \mathit{Normal}(\mu, \Sigma_1) \qquad \mathbf{y}|\mathbf{x} \sim \mathit{Normal}(\mathbf{A}\mathbf{x} + \mathbf{b}, \Sigma_2)$$

$$\Rightarrow \quad \mathbf{x}|\mathbf{y} \sim \mathit{Normal}(\overline{\mu}, \overline{\Sigma})$$

$$\overline{\mu} = (\Sigma_1^{-1} + \mathbf{A}^T\Sigma_2^{-1}\mathbf{A})^{-1}(\mathbf{A}^T\Sigma_2^{-1}(\mathbf{y} - \mathbf{b}) + \Sigma_1^{-1}\mu)$$
$$\overline{\Sigma} = (\Sigma_1^{-1} + \mathbf{A}^T\Sigma_2^{-1}\mathbf{A})^{-1}$$

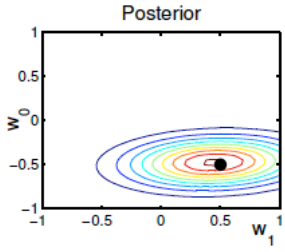
# Regressione lineare
# //Approccio Bayesiano: l'ipotesi sul prior

- Nel nostro caso, con un po' di conti... (e non scrivendo le variabili x per semplificare la notazione)

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

$$\Rightarrow \quad p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w})$$

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\Phi^{\mathrm{T}}\mathbf{t})$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\Phi^{\mathrm{T}}\Phi$$




- con $\mathbf{m}_0 = 0$ $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$ $\Rightarrow$ $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$

$$\mathbf{m}_N = \beta\mathbf{S}_N\Phi^{\mathrm{T}}\mathbf{t}$$
$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\Phi^{\mathrm{T}}\Phi$$

# Regressione lineare //Approccio Bayesiano: l'ipotesi sul prior

- In questo setting

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$$
$$\mathbf{m}_N = \beta \mathbf{S}_N \Phi^{\mathrm{T}} \mathbf{t}$$
$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \Phi^{\mathrm{T}} \Phi$$

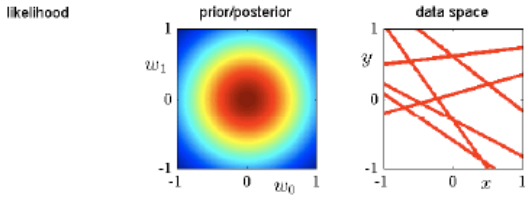
- Si noti che
  - 1. Per $\alpha \to 0$ ⇒ $\mathbf{m}_N \to \mathbf{w}_{\mathrm{ML}} = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\mathbf{t}$
    - ritroviamo la regressione di ML
  - 2. Calcolando la log posteriori

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\right)^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const}$$

    - ritroviamo la ML regolarizzata

# Regressione lineare //Approccio Bayesiano: learning sequenziale

$$y(x, w_0, w_1) = w_0 + w_1 x$$

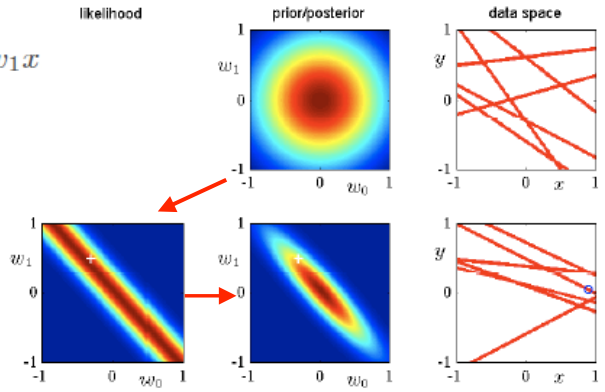



$$p(w_0, w_1)$$
$$\mu = 0 \text{ e } \Sigma = \sigma^2 \mathbf{I} = 0.04\mathbf{I}$$

rette campionate dalla distribuzione

# Regressione lineare
# //Approccio Bayesiano: learning sequenziale

$y(x, w_0, w_1) = w_0 + w_1 x$




$p(w_0, w_1)$

$\mu = \mathbf{0}$ e $\Sigma = \sigma^2 \mathbf{I} = 0.04\mathbf{I}$

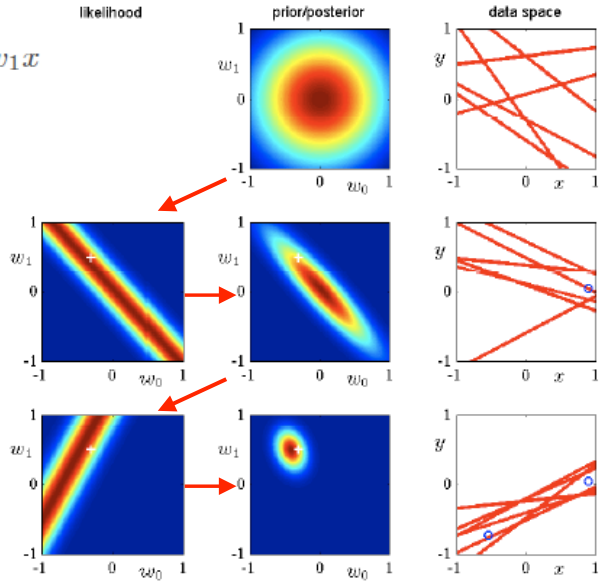
rette campionate dalla distribuzione

osservo $(x_1, y_1)$

$p(w_0, w_1 | x_1, y_1)$

# Regressione lineare
# //Approccio Bayesiano: learning sequenziale

$y(x, w_0, w_1) = w_0 + w_1 x$




$p(w_0, w_1)$

$\mu = \mathbf{0}$ e $\Sigma = \sigma^2 \mathbf{I} = 0.04\mathbf{I}$

rette campionate dalla distribuzione

osservo $(x_1, y_1)$

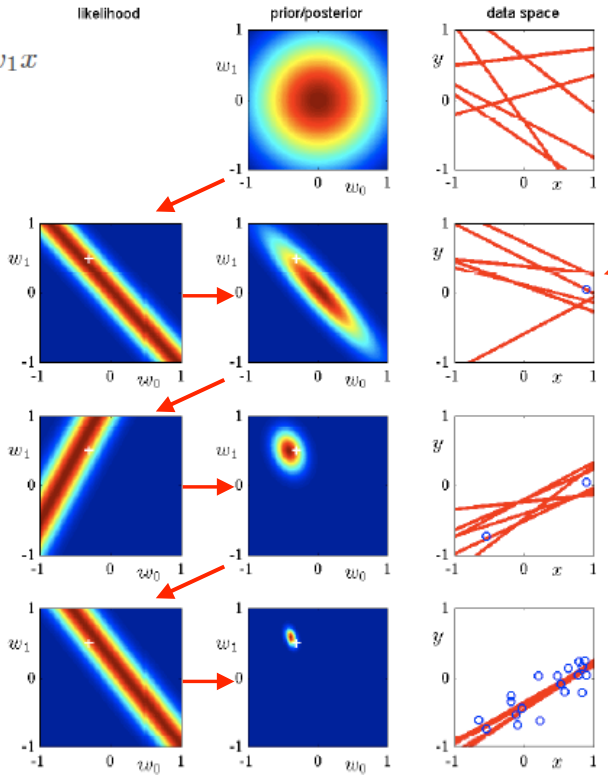
$p(w_0, w_1 | x_1, y_1)$

osservo $(x_2, y_2)$

$p(w_0, w_1 | x_1, y_1, x_2, y_2)$

# Regressione lineare
# //Approccio Bayesiano: learning sequenziale

$y(x, w_0, w_1) = w_0 + w_1 x$




$p(w_0, w_1)$

$\mu = \mathbf{0}$ e $\Sigma = \sigma^2 \mathbf{I} = 0.04\mathbf{I}$

rette campionate dalla distribuzione

osservo $(x_1, y_1)$

$p(w_0, w_1 | x_1, y_1)$

osservo $(x_2, y_2)$

$p(w_0, w_1 | x_1, y_1, x_2, y_2)$

$p(w_0, w_1 | x_1, y_1, \ldots, x_n, y_n)$

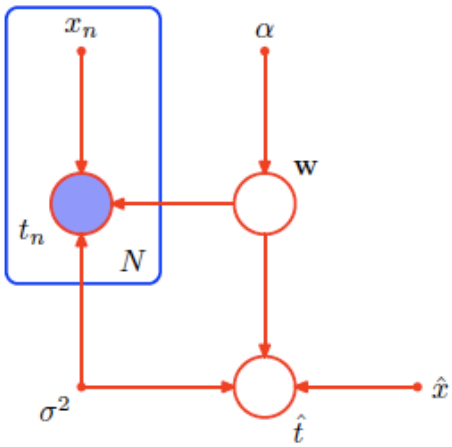
# Regressione lineare
# //Approccio Bayesiano: predizione

- Data la posteriori $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) \times p(\mathbf{w}|\alpha)$

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta) d\mathbf{w}$$

$$p(t|\mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

$$p(\mathbf{w}|\mathbf{t}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$




⇒ L'integrale rappresenta la convoluzione di due Gaussiane

⇒ Risultato ancora gaussiano

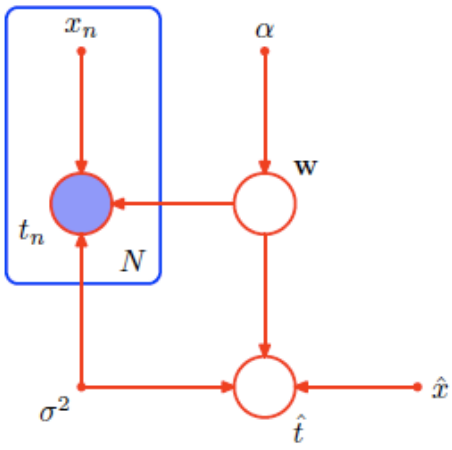
$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}} \boldsymbol{\Phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

$$\sigma_N^2(\mathbf{x}) = \underbrace{\beta^{-1}}_{\text{Rumore dati}} + \underbrace{\boldsymbol{\Phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\Phi}(\mathbf{x})}_{\text{Incertezza nei parametri } \mathbf{w}}$$

# Regressione lineare //Approccio Bayesiano: predizione

$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\Phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

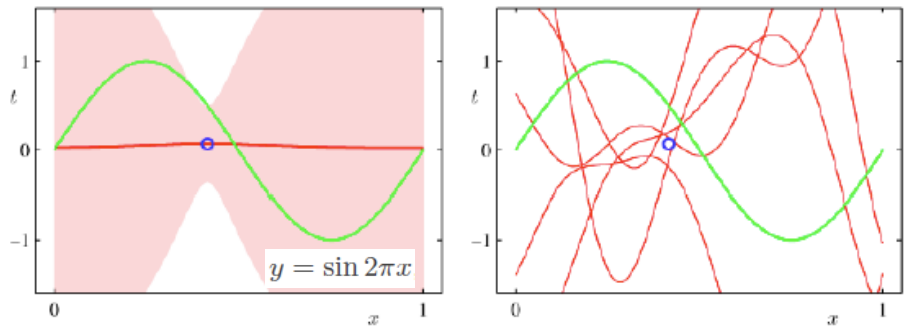
$$\sigma_N^2(\mathbf{x}) = \underbrace{\beta^{-1}}_{\text{Rumore dati}} + \underbrace{\boldsymbol{\Phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\Phi}(\mathbf{x})}_{\text{Incertezza nei parametri } \mathbf{w}}$$



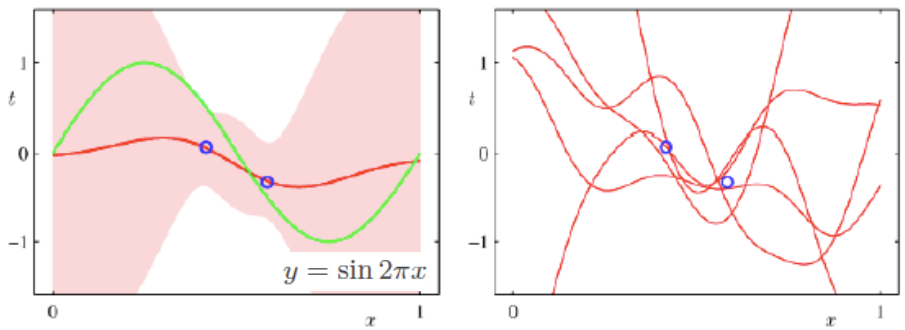

- Al crescere del numero di elementi nel training set:
  - il secondo termine della varianza diminuisce, e quindi la distribuzione tende a concentrarsi intorno al valore previsto
  - solo il primo termine rimane significativo, mostrando che la sola incertezza rimanente è quella relativa ai dati osservati

# Regressione lineare //Approccio Bayesiano: esempio

1 punto nel training set

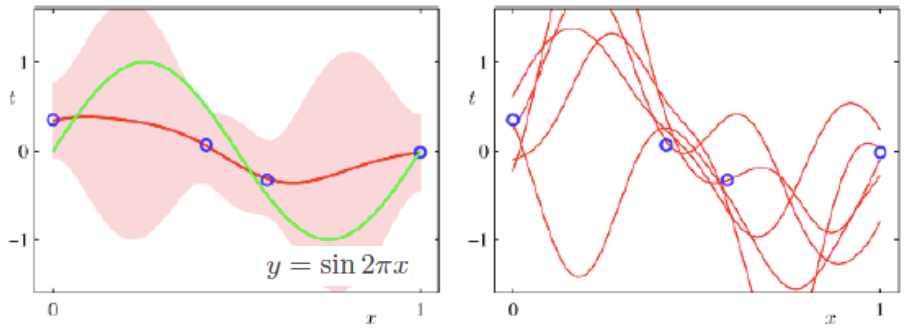



2 punti nel training set

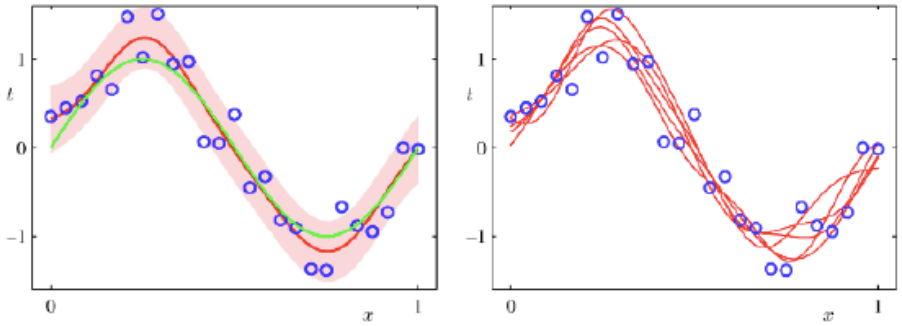


# Regressione lineare
# //Approccio Bayesiano: esempio

4 punti nel training set




25 punti nel training set



# Esempio Matlab

$$f(x_n;\mathbf{w}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}_n$$

$$w_0 \sim \mathcal{N}(0,\alpha) \qquad w_1 \sim \mathcal{N}(0,\alpha)$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}_{w_0}(0,\alpha)\mathcal{N}_{w_1}(0,\alpha) = \mathcal{N}_{\mathbf{w}}(\mathbf{0},\mathbf{\Lambda}) \qquad \mathbf{\Lambda} = \alpha\mathbf{I}$$

$$\prod_{n=1}^{N}\mathcal{N}_{t_n}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n,\sigma) = \mathcal{N}_{\mathbf{t}}(\mathbf{X}\mathbf{w},\sigma\mathbf{I})$$

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha) = \frac{\mathcal{N}_{\mathbf{t}}(\mathbf{X}\mathbf{w},\sigma\mathbf{I})\mathcal{N}_{\mathbf{w}}(\mathbf{0},\mathbf{\Lambda})}{\int \mathcal{N}_{\mathbf{t}}(\mathbf{X}\mathbf{w},\sigma\mathbf{I})\mathcal{N}_{\mathbf{w}}(\mathbf{0},\mathbf{\Lambda})d\mathbf{w}} = \mathcal{N}(\boldsymbol{\mu},\mathbf{\Sigma})$$

$$\boldsymbol{\mu} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}$$

$$\mathbf{\Sigma} = \sigma^2\left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}$$

# Esempio Matlab

$$f(x_n; \mathbf{w}) = \mathbf{w}^\mathsf{T}\mathbf{x}_n$$

$$\begin{aligned} E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}|\mathbf{x}_{new}\right\} &= E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{\mathbf{x}_{new}^\mathsf{T}\mathbf{w}\right\} \\ &= \mathbf{x}_{new}^\mathsf{T}\int \mathbf{w}p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)d\mathbf{w} \\ &= \mathbf{x}_{new}^\mathsf{T}\boldsymbol{\mu} = \mathbf{x}_{new}^\mathsf{T}\left(\mathbf{X}^\mathsf{T}\mathbf{X}+\frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \end{aligned}$$

$$\begin{aligned} \text{var}(t_{new}|\mathbf{x}_{new}) &= E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}^2|\mathbf{x}_{new}\right\} - E^2_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}|\mathbf{x}_{new}\right\} \\ &= \mathbf{x}_{new}^\mathsf{T}E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma)}\left\{\mathbf{w}\mathbf{w}^\mathsf{T}\right\}\mathbf{x}_{new} - \left(\mathbf{x}_{new}^\mathsf{T}E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma)}\left\{\mathbf{w}\right\}\right)^2 \\ &= \mathbf{x}_{new}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{x}_{new} = \sigma^2\mathbf{x}_{new}^\mathsf{T}\left(\mathbf{X}^\mathsf{T}\mathbf{X}+\frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{x}_{new} \end{aligned}$$

# Esempio Matlab

N=25; xx=rand(N,1)-0.5;
X=[xx.^2 xx]; w0=-0.5; w1=0.5; sigma=0.05; alpha = 1;

f = X*[w1; w0];

$$f(x_n; \mathbf{w}) = \mathbf{w}^\mathsf{T}\mathbf{x}_n$$

simulo dati di training

t = f+sigma*randn(N,1);

$$t|x \sim \mathcal{N}(f(x;\mathbf{w}),\sigma)$$

that = X*inv(X'*X + eye(2)*sigma/alpha)*X'*t;

$$E_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma,\alpha)}\left\{t_{new}|\mathbf{x}_{new}\right\} = \mathbf{x}_{new}^\mathsf{T}\left(\mathbf{X}^\mathsf{T}\mathbf{X}+\frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

predizione

[x,y]=meshgrid(-1:0.05:1,-1:0.05:1);
[n,n]=size(x);

W=[reshape(x, n*n, 1) reshape(y, n*n, 1)];

mu = inv(X'*X + eye(2)*sigma/alpha)*X'*t;

$$\boldsymbol{\mu} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}+\frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$

C = sigma*inv(X'*X + eye(2)*sigma/alpha);

$$\boldsymbol{\Sigma} = \sigma^2\left(\mathbf{X}^\mathsf{T}\mathbf{X}+\frac{\sigma^2}{\alpha}\mathbf{I}\right)^{-1}$$

prior = reshape(gauss([0 0], eye(2)*alpha, W),[n n]);

$$\mathcal{N}_\mathbf{w}(\mathbf{0},\boldsymbol{\Lambda}) \qquad \boldsymbol{\Lambda} = \alpha\mathbf{I}$$

likelihood = reshape(gauss(t',eye(N)*sigma,W*X'),[n n]);

$$\prod_{n=1}^N \mathcal{N}_{t_n}(\mathbf{w}^\mathsf{T}\mathbf{x}_n,\sigma)$$

posterior = reshape(gauss(mu',C, W),[n n]);

$$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$$

calcola le distribuzioni per plottarle

# Esempio Matlab

```
figure

subplot(2,2,1)
contour(x,y,prior)
hold
plot(w1,w0,'o')

subplot(2,2,2)
contour(x,y,likelihood)
hold
plot(w1,w0,'o')

subplot(2,2,3)
contour(x,y,posterior);
hold
plot(w1,w0,'o')

subplot(2,2,4)
hold
[X_sort,ind_sort]=sort(xx(:,1));
plot(xx(ind_sort,1),f(ind_sort));
plot(xx(ind_sort,1),t(ind_sort),'.');
plot(xx(ind_sort,1),that(ind_sort),'r');
```