Machine Intelligence 2

6.1 Maximum Likelihood & Estimation Theory

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

SS 2017

Estimation theory

Estimator

An estimator $\hat{P}(X)$ is a function that maps from its sample space X (data) to a set of *sample estimates* W

An estimator ...

- is a function of a random variable
- is a random variable
- can be statistically characterized via its moments (mean, variance, ...)
  $\rightsquigarrow$ quality criteria: unbiasedness, efficiency

Probability distributions: an example

$$P\Big(\big\{\underline{\mathbf{x}}^{(\alpha)}\big\};\underline{\mathbf{w}}^*\Big)$$

set of observations: $\big\{\underline{\mathbf{x}}^{(\alpha)}\big\}, \alpha = 1, \dots, p$ from true distribution

Goal: estimate "true" values $\underline{\mathbf{w}}^*$ from observed data

estimator $\underline{\widehat{\mathbf{w}}}$:

$$\underline{\widehat{\mathbf{w}}} = \underline{\widehat{\mathbf{w}}}\big(\big\{\underline{\mathbf{x}}^{(\alpha)}\big\}\big)$$

- procedure for the determination of $\underline{\mathbf{w}}^*$ given the observed data
- $\underline{\mathbf{w}}^*$ is a function of $\big(\big\{\underline{\mathbf{x}}^{(\alpha)}\big\}\big)$
- $\underline{\mathbf{x}}^{(\alpha)}$ are random variables $\rightarrow \underline{\widehat{\mathbf{w}}}$ is a **random variable!**

The Maximum Likelihood estimator

the likelihood function

$$\widehat{P}\big(\{\underline{\mathbf{x}}^{(\alpha)}\};\underline{\mathbf{w}}\big)$$

the $\log$**-likelihood function**

$$\ln \widehat{P}\big(\{\underline{\mathbf{x}}^{(\alpha)}\};\underline{\mathbf{w}}\big) = \sum_{\alpha=1}^{p} \ln \widehat{P}\big(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}}\big)$$

the Maximum Likelihood estimator

$$\widehat{\underline{\mathbf{w}}} = \underset{\underline{\mathbf{w}}}{\operatorname{argmax}} \widehat{P}\big(\{\underline{\mathbf{x}}^{(\alpha)}\};\underline{\mathbf{w}}\big)$$

Quality criteria for estimators

What are good estimators?

bias: $$\underline{\mathbf{b}} = \underbrace{\left\langle \widehat{\underline{\mathbf{w}}} \right\rangle_{P(x^\alpha;w)}}_{\substack{\text{expectation} \\ \text{w.r.t the \underline{true}} \\ \text{distribution}}} - \underline{\mathbf{w}}^*$$

variance: $$\underline{\boldsymbol{\Sigma}} = \left\langle (\widehat{\underline{\mathbf{w}}} - < \hat{\underline{\mathbf{w}}} >)(\widehat{\underline{\mathbf{w}}} - < \hat{\underline{\mathbf{w}}} >)^T \right\rangle_{P(x^\alpha;w)}$$

Optimal estimators

no bias: $$\underline{\mathbf{b}} \stackrel{!}{=} 0$$ $\leftarrow$ only possible if true model within model class

minimal variance: $$|\underline{\boldsymbol{\Sigma}}| \stackrel{!}{=} \min$$

The sample mean

N observations $x^{(\alpha)}$

$$x^{(\alpha)} = A + \epsilon^{(\alpha)}$$

with $\epsilon^{(\alpha)} \sim N(0, \sigma^2)$

Examples for estimators for A:

$\hat{A} = \frac{1}{N} \sum x^{(\alpha)}$	unbiased
$\tilde{A} = \frac{1}{2N} \sum x^{(\alpha)}$	biased for $A \neq 0$
$\tilde{A} = k$	minimum variance but biased

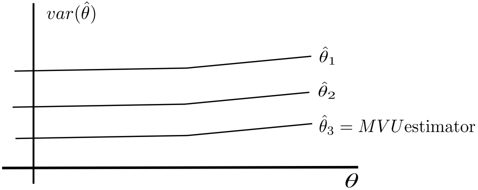
The Minimum Variance Unbiased estimator

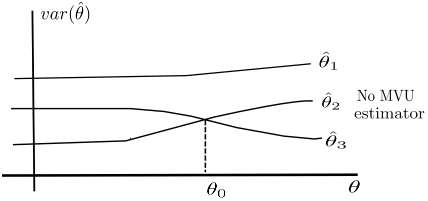
Optimal estimators

no bias: $\underline{\mathbf{b}} \stackrel{!}{=} 0$ $\leftarrow$ only possible if true model within model class

minimal variance: $|\underline{\mathbf{\Sigma}}| \stackrel{!}{=} \min$

MVU: criteria have to hold for ALL possible values of $\underline{\mathbf{w}}^*$!





MVUs do not always exist

The Minimum Variance Unbiased estimator

given just observed sample conditionally independent observations with the 2 pdfs

$$x[0] \sim \mathcal{N}(\theta, 1) \qquad x[1] \sim \begin{cases} \mathcal{N}(\theta, 1) & \text{if } \theta \geq 0 \\ \mathcal{N}(\theta, 2) & \text{if } \theta < 0 \end{cases}$$

two estimators

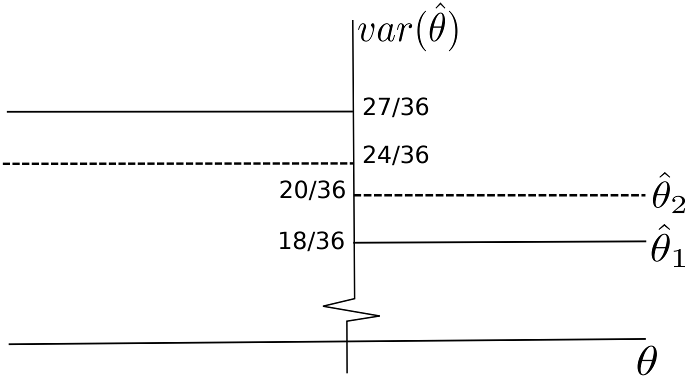
$$\hat{\theta}_1 = \frac{1}{2}(x[0] + x[1]) \qquad \text{and} \qquad \hat{\theta}_2 = \frac{2}{3}x[0] + \frac{1}{3}x[1]$$

variances:

$$var(\hat{\theta}_1) = \frac{1}{4}(var(x[0]) + var(x[1])) \qquad \begin{cases} \frac{18}{36} & \text{if } \theta \geq 0 \\ \frac{27}{36} & \text{if } \theta < 0 \end{cases}$$

$$var(\hat{\theta}_2) = \frac{4}{9}var(x[0]) + \frac{1}{9}var(x[1]) \qquad \begin{cases} \frac{20}{36} & \text{if } \theta \geq 0 \\ \frac{24}{36} & \text{if } \theta < 0 \end{cases}$$

Example for the non-existence of MVUs (Kay, 1993)



MVU vs. minimal mean squared error

$$MSE(\underline{\hat{\mathbf{w}}}) = E[(\underline{\hat{\mathbf{w}}} - \underline{\mathbf{w}^*})^2]$$

This however does not yield a realizable estimator because

$$\begin{aligned}
MSE(\hat{w}) &= E\left\{[(\hat{w} - E(\hat{w})) + (E(\hat{w}) - w^*)]^2\right\} \\
&= var(\hat{w}) + [E(\hat{w}) - w^*]^2 \\
&= variance + bias^2
\end{aligned}$$

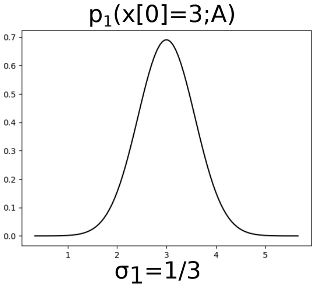
MSE trades bias against variance.

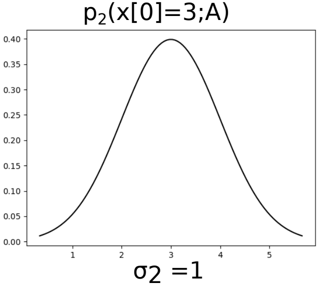
Cramer-Rao bound for unbiased estimators

The stronger a PDF depends on its parameters, the more accurate will their estimates be.

N observations $x^{(\alpha)}$ with $\epsilon^{(\alpha)}\ N(0, \sigma^2)$

$$x^{(\alpha)} = A + \epsilon^{(\alpha)}, \qquad \hat{A} = \frac{1}{N} \sum x^{(\alpha)}$$





Accuracy can be measured by the 'sharpness' of the likelihood function ($\rightsquigarrow$ 2nd derivative of the neg. log likelihood).

Cramer-Rao bound for unbiased estimators

Fisher information matrix (Hessian matrix):

$$H_{ij} = - \left\langle \frac{\partial^2 \ln P}{\partial \mathrm{w}_i \partial \mathrm{w}_j} \right\rangle_{P(x^\alpha; w*)} \Bigg|_{\underline{\mathbf{w}}^*}$$

For all unbiased estimators the following holds (Cramer-Rao Bound):

$$\underline{\mathbf{\Sigma}} - \left(\underline{\mathbf{H}}^{-1} \right) \text{ is a positive semidefinite matrix}$$

it follows:

$$var(\hat{w}_i) \geq [H^{-1}]_{ii} \text{ for all } i$$

Variance of an estimator > 1/ Fisher Information

This is a <u>universal</u> lower bound on the variance of estimators. The bound is tight.

Example: CRB for a scalar parameter w

The property of "positive semidefinite":

$$\sigma_{\mathrm{w}}^2 - \left\{ - \left\langle \frac{d^2 \ln P}{d\mathrm{w}^2} \right\rangle_p \bigg|_{\underline{\mathbf{w}}^*} \right\}^{-1} \geq 0$$

$$\sigma_{\mathrm{w}}^2 > -\frac{1}{\left\langle \frac{d^2 \ln P}{d\mathrm{w}^2} \right\rangle_p \bigg|_{\underline{\mathbf{w}}^*}}$$

Comment

Fisher information: precision of the estimator / interesting measure for evaluating data representations

Good estimators

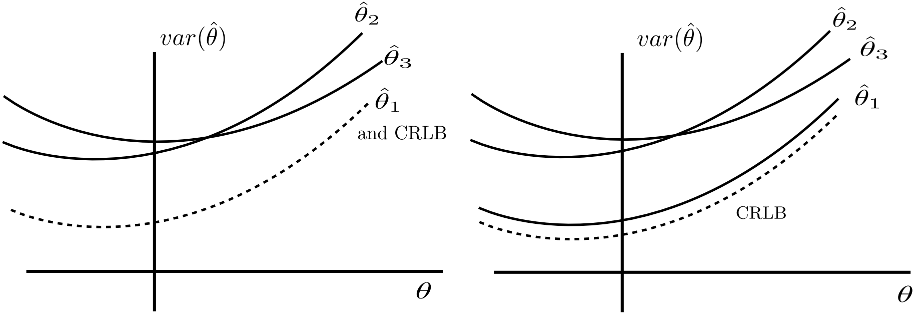
efficient estimator:

$$\underline{\mathbf{b}} = \underline{\mathbf{0}} \text{ and} \underline{\mathbf{\Sigma}} = \underline{\mathbf{H}}^{-1} \qquad \leftarrow \text{variance assumes lower bound}$$

unbiased minimum variance estimator:

$$\underline{\mathbf{b}} = \underline{\mathbf{0}} \text{ and} \left| \underline{\mathbf{\Sigma}} - \underline{\mathbf{H}}^{-1} \right| \stackrel{!}{=} \min_{\text{all estimators}}$$

Illustration: Cramer-Rao bound



Asymptotic optimality

An estimator is said to be asymptotically unbiased if for $p \to \infty$(limit of infinite sample size):

$$E(\underline{\hat{\mathbf{w}}}) \to \underline{\mathbf{w}}^*$$

An estimator is said to be asymptotically efficient if for $p \to \infty$:

$$var(\underline{\hat{\mathbf{w}}}) \to \text{Cramer Rao lower bound}$$

An estimator is said to be consistent if it converges to the true value for $p \to \infty$ and is asymptotically unbiased.

Results for the Maximum Likelihood estimator

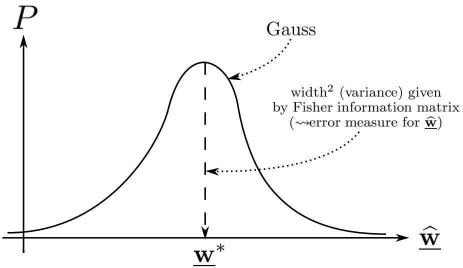
$P\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\};\underline{\mathbf{w}}\right)$ normalized and two times differentiable

$H_{ij} = -\left\langle \frac{\partial^2 \ln P}{\partial \mathrm{w}_i \partial \mathrm{w}_j} \right\rangle_{P(x^\alpha;w*)}$ Fisher information matrix

The Maximum Likelihood estimator is consistent and asympotically unbiased and efficient.

$$\widehat{\underline{\mathbf{w}}} \sim \mathcal{N}\left(\underline{\mathbf{w}}^*, \underline{\mathbf{H}}^{-1}_{(\underline{\mathbf{w}}^*)}\right)$$

asymptotically Gaussian distributed



Summary

- An estimator is a random variable.
- $\Rightarrow$ It can only be analyzed statistically (e.g. mean, variance, shape of distribution).
- biased & unbiased estimators
- Minimum Variance Unbiased estimator (MVU) has smallest variance for **all values** of the true parameter

MVUs and the Cramer-Rao bound

- Minimum Variance Unbiased estimators do not always exist
- Cramer Rao Bound provides a universal bound but may not be realizable

Outlook

Inclusion of prior knowledge

- MLEs: no prior knowledge regarding 'reasonable' parameter values
- Maximum A Posteriori estimates (MAP) incorporate such knowledge via Bayes Theorem ($\rightsquigarrow$ regularisation)

$$p(\underline{\mathbf{w}}|\underline{\mathbf{x}}) \propto p(\underline{\mathbf{x}}|\underline{\mathbf{w}})p(\underline{\mathbf{w}})$$

- Beyond point estimates: Bayesian statistics. A complete (probabilistic) treatment should exploit the degrees of belief in a given model (set of parameters)

Model Fitting: Bayes & Maximimum Likelihood

Estimators revisited

set of observations: $\left\{x^{(\alpha)}\right\}, \alpha = 1, \ldots, p$

$$P\Big(\left\{x^{(\alpha)}\right\}; \underline{\mathbf{w}}^*\Big)$$

drawn from the true distribution: $x^{(\alpha)} \in \{1, 2\}$

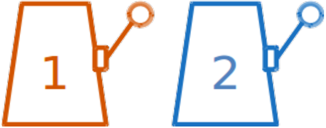
Goal: estimate "true" values $\underline{\mathbf{w}}^*$ from observed data

estimator $\widehat{\underline{\mathbf{w}}}$:

$$\widehat{\underline{\mathbf{w}}} = \widehat{\underline{\mathbf{w}}}\big(\left\{x^{(\alpha)}\right\}\big)$$

- procedure for the determination of $\underline{\mathbf{w}}^*$ given the observed data
- $\underline{\mathbf{w}}^*$ is a function of $\big(\left\{x^{(\alpha)}\right\}\big)$
- $x^{(\alpha)}$ are random variables $\rightarrow \widehat{\underline{\mathbf{w}}}$ is a **random variable!**

A two-armed bandit task



$$\mathcal{R}(x=1) = \begin{cases} 1, p = 0.7 \\ 0, p = 0.3 \end{cases}$$

$$\mathcal{R}(x=2) = \begin{cases} 1, p = 0.3 \\ 0, p = 0.7 \end{cases}$$

$$Q^{(\alpha+1)}(x^{(\alpha)}) = Q^{(\alpha)}(x^{(\alpha)}) + w^{(1)}(R^{(\alpha)}(x^{(\alpha)}) - Q^{(\alpha)}(x^{(\alpha)})$$

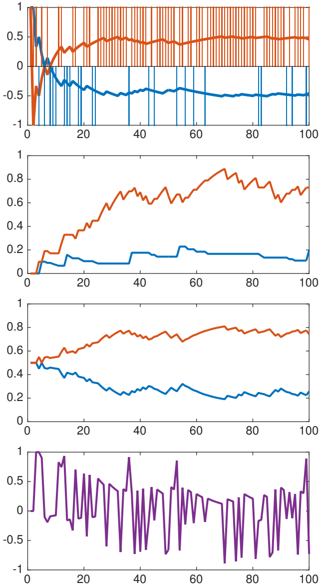
$$p^{(\alpha)}(x^{(\alpha)}) = \frac{e^{w^{(2)}}Q(x^{(\alpha)})}{\sum_i e^{w^{(2)}}Q(x_i)}$$

$$P\Big(\{\underline{\mathbf{x}}\}; \underline{\mathbf{w}}^*\Big) = \prod_{\alpha=1}^{p} p^{(\alpha)}(x^{(\alpha)})$$

"true" parameters:

$$w^{(1)} = 0.1, w^{(2)} = 2, Q^{(1)}(x=1) = Q^{(1)}(x=2) = 0$$

Agents playing games



R

$x = 1$ $\quad$ $x = 2$

average (per trial) cumulative reward from the actions 1 and 2.

Q

corresponding Q values

$p(x)$

probability of actions 1 and 2

prediction error

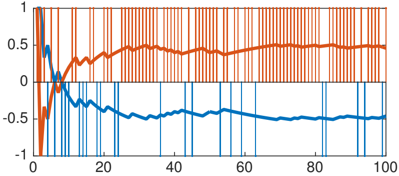
reward prediction error for the chosen action

$\delta^{(\alpha)} = R^{(\alpha)}(x^{(\alpha)}) - Q^{(\alpha)}(x^{(\alpha)})$

Model fitting

Data generated using the "true" paramters

$$w^{(1)} = 0.1, w^{(2)} = 2, Q^{(1)}(x=1) = Q^{(1)}(x=2) = 0$$

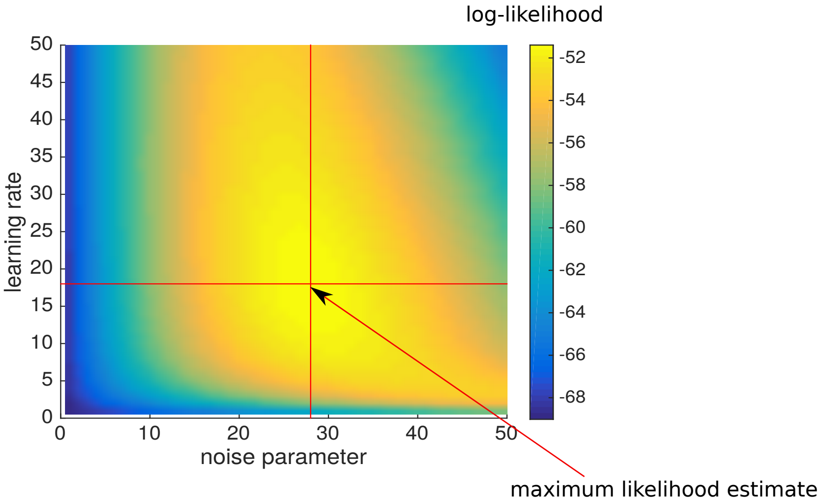


- compute the corresponding log-likelihood function:

$$\mathcal{L}(\{\underline{\mathbf{x}}\}; \underline{\widehat{\mathbf{w}}}) = \ln(\prod_{\alpha=1}^{p} p^{(\alpha)}(x^{(\alpha)})) = \sum_{\alpha=1}^{p} (w^{(1)} Q(x^{(\alpha)}) - \ln(\sum_{i} e^{w^{(2)} Q(x_i)})$$

- choose $\underline{\widehat{\mathbf{w}}}$ which correspond to the maximum of the log-likelihood function.

Model fitting: grid search



"true" paramters $w^{(1)} = 0.1, w^{(2)} = 2, Q^{(1)}(x=1) = Q^{(1)}(x=2) = 0$

Model comparison

To avoid overfitting, we compare models according to the model evidence

$$P(\underline{\mathbf{x}}|M) = \int P(\underline{\mathbf{x}}|M, \underline{\mathbf{w}})P(\underline{\mathbf{w}}|M)d\underline{\mathbf{w}}$$

which requires computing very high-dimensional integral and analytically intractable posterior distributions.

We can use a Gaussian distribution to approximate $P(\underline{\mathbf{x}}|M, \underline{\mathbf{w}})P(\underline{\mathbf{w}}|M) := f(\underline{\mathbf{w}})$ around its mode $\hat{\underline{\mathbf{w}}}_{MAP}$:

$$\int f(\underline{\mathbf{w}})d\underline{\mathbf{w}} \approx f(\hat{\underline{\mathbf{w}}}_{MAP}) \int \exp(-\frac{1}{2}(\underline{\mathbf{w}} - \hat{\underline{\mathbf{w}}}_{MAP})^T H(\underline{\mathbf{w}} - \hat{\underline{\mathbf{w}}}_{MAP}))d\underline{\mathbf{w}} = f(\hat{\underline{\mathbf{w}}}_{MAP}) 2\pi^{\frac{n}{2}} H^{\frac{-1}{2}}$$

Laplace approximation

$$\ln(P(\underline{\mathbf{x}}|M)) \approx \underbrace{\ln(P(\underline{\mathbf{x}}|M, \hat{\underline{\mathbf{w}}}_{MAP}))}_{\text{log likelihood at the optimized parameters}} + \underbrace{\ln(P(\hat{\underline{\mathbf{w}}}_{MAP}|M)) + \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|H|)}_{\text{penalizes model complexity}}$$

Model comparison

- Bayesian Information Criterion
  - → BIC simplifies Laplace approximation by assuming that the sample size is large and retains terms which grow with the number of data points only.
  - → Prior over parameters: Gaussian with broad variance (large sample size); data: iid distributed $\ln(|H|) \approx n \ln(p)$. Only terms $\mathcal{O}(\ln(p))$ are retained.

$$BIC \approx \ln(P(\underline{\mathbf{x}}|M, \hat{\underline{\mathbf{w}}}_M)) - \frac{n}{2} \ln(p)$$

  n =number of free parameters

- other penalized scores for model comparison:
  - → $AIC = \ln(P(\underline{\mathbf{x}}|M, \hat{\underline{\mathbf{w}}}_M)) - n$, penalizes the number of parameters less strongly than does BIC
  - → $DIC = D(\bar{\underline{\mathbf{w}}}) + 2p_D$, a hierarchical modeling generalization of the Bayesian information criteria, model complexity measured by estimate of the effective number of parameters.
  - → WAIC, LOO...

Model comparison

- Data generated using the "true" paramters $w^{(1)} = 0.1, w^{(2)} = 2, Q^{(1)}(x = 1) = Q^{(1)}(x = 2) = 0$
- model 1: 2 free parameters: $w^{(1)}, w^{(2)}$.
- model 2: 3 free parameters: $w^{(1)}, w^{(2)}, w^{(3)} = Q^{(1)}(x = 1) = Q^{(1)}(x = 2)$
- maximum likelihood estimate (grid search):

$$\text{Model1: } \hat{w}^1 = 0.009, \hat{w}^2 = 2.8$$
$$\text{Model2: } \hat{w}^1 = 0.009, \hat{w}^2 = 4.9, \hat{w}^3 = 2.8$$

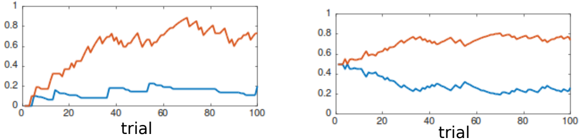
- model comparison by BIC scores:

$$\text{BIC}_1 = -51.36 - \frac{2}{2}\ln(100) = -55.97$$
$$\text{BIC}_2 = -51.42 - \frac{3}{2}\ln(100) = -58.33$$
$$\text{BIC}_{random} = 100\ln(0.5) \qquad = -69.32$$

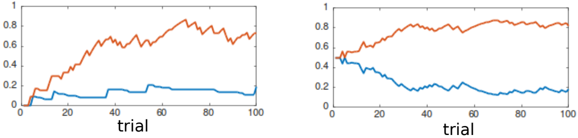
Model performance

$$Q^{(\alpha+1)}(x^{(\alpha)}) = Q^{(\alpha)}(x^{(\alpha)}) + w^{(1)}(R^{(\alpha)}(x^{(\alpha)}) - Q^{(\alpha)}(x^{(\alpha)})) \qquad p^{(\alpha)}(x^{(\alpha)}) = \frac{e^{w^{(2)}} Q(x^{(\alpha)})}{\sum_i e^{w^{(2)}} Q(x_i)}$$

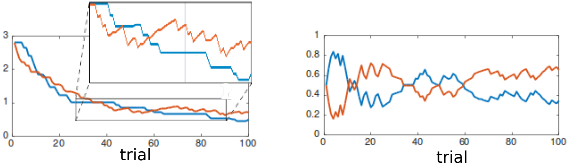
Reference agent



Model 1



Model 2



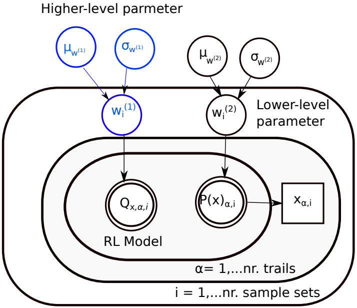
Model fitting: Maximum Likelihood estimation

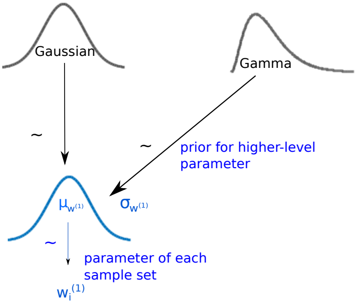
- nonlinear function optimization: given a function to compute the likelihood with some free parameters.
- local search: likelihood surfaces may not be well- behaved but may have multiple peaks. The optimization may run into local minimum.
- model fitting process is feasible but finicky, which requires ongoing monitoring and tuning.

Choice of priors

- prior information: the likely range of the parameters via $P(\underline{\mathbf{w}}_M|M)$. Adopting a model with parameter priors (typically Gaussian or Beta distributions) gives us a two-level hierarchical model of how a full dataset is produced.
- probability distributions (from each level of the hieararchical model) can be approximated by drawing independent samples: Markov Chain Monte Carlo methods.
- MCMC method: approximating a distribution with a large set of samples and each sample is drawn based on the previous sample

Model fitting: Hierarchical Bayesian analysis





- assume the lower-level parameters come from a Gaussian prior, we can estimate parameters of the higher-level $(\mu_{w^{(1)}}, \sigma_{w^{(1)}}, \mu_{w^{(2)}}, \sigma_{w^{(2)}})$ to see the group differences.

$$P(\underline{\mathbf{x}}_i|\mu_{w^{(1)}}, \sigma_{w^{(1)}}, \mu_{w^{(2)}}, \sigma_{w^{(2)}}) =$$

$$\int P(\underline{\mathbf{x}}_i|w_i^{(1)}, w_i^{(2)}) P(w_i^{(1)}|\mu_{w^{(1)}}, \sigma_{w^{(1)}}) P(w_i^{(2)}|\mu_{w^{(2)}}, \sigma_{w^{(2)}}) dw_i^{(1)} dw_i^{(2)}$$