Wiederholung: Von Bayes‘schem Schließen zu Maximum Likelihood

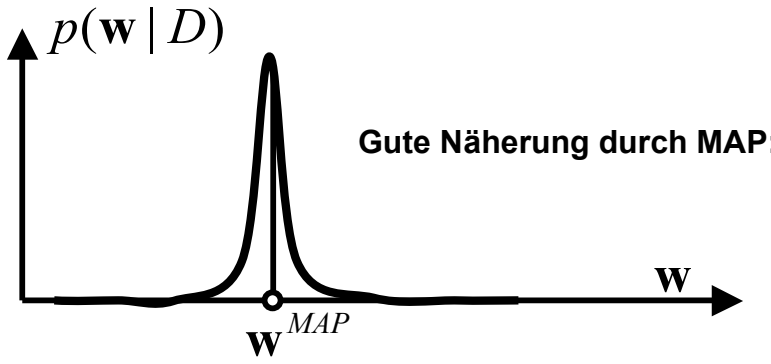
- **Bayes‘sche Datenmodellierung:**

$$p(\mathbf{w} \mid D) = \frac{1}{p(D)} p(D \mid \mathbf{w}) p(\mathbf{w})$$

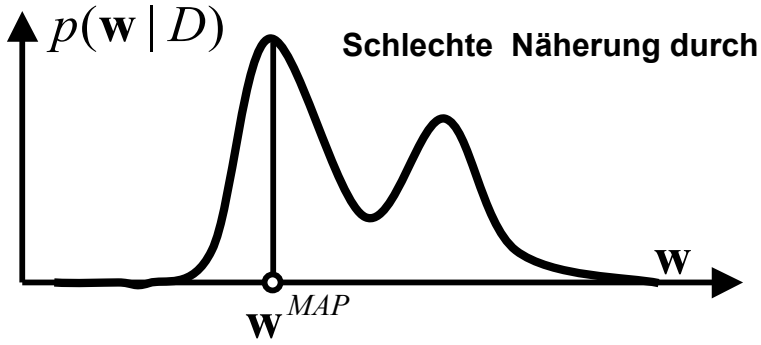
- **MAP-Näherung: Verwende das Modell mit der größten a-posteriori-Wahrscheinlichkeit**

$$\mathbf{w}^{MAP} = \arg\max_{\mathbf{w}} p(\mathbf{w} \mid D)$$

- **I. d. Regel gute Näherung bei vielen Daten**
- **Übergang zu frequentistischer Sicht (relative Häufigkeiten, ein wahres Modell)**



Gute Näherung durch MAP:



Schlechte Näherung durch MAP:

- **Minimiere negativen log posterior:**

$$-\ln p(D \mid \mathbf{w}) - \ln p(\mathbf{w}) + const = \min$$

⇧

Maximum Likelihood (=min neg. log. likelihood)

Später: **Fehlerminimierung**

Maximum-Likelihood-Parameterschätzung

Ziel:

- **Maximiere Wahrscheinlichkeit, dass die Daten aus dem Modell erzeugt wurden => Maximum Likelihood**
- **Bem: Maximum-Likelihood (ML) entspricht MAP ohne Vorwissen**

Prinzip:

- **Likelihood:** $p(D \mid \mathbf{w}) = p(\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}\} \mid \mathbf{w})$
- **Max. Likelihood:** $\mathbf{w}^{ML} = \arg\max_{\mathbf{w}} p(D \mid \mathbf{w})$
- **Äquivalent:** $\mathbf{w}^{ML} = \arg\min_{\mathbf{w}} -\ln p(\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}\} \mid \mathbf{w}) =: \arg\min_{\mathbf{w}} L(\mathbf{w})$

  $L(\mathbf{w})$ **wird auch als negative log-likelihood oder „Likelihood-Funktion“ bezeichnet**

- **Für iid („independent, identically distributed“) gezogene Datenpunkte sind die Likelihood-Funktionen verschiedener Datenpunkte unabhängig:**

$$L(\mathbf{w}) = -\ln p\left(\left\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(M)}\right\} \mid \mathbf{w}\right) = -\ln \prod_{m=1}^{M} p(\mathbf{x}^{(m)} \mid \mathbf{w}) = -\sum_{m=1}^{M} \ln p(\mathbf{x}^{(m)} \mid \mathbf{w})$$

Lernen von Datenmodellen

Maximum-Likelihood Schätzung und Fehlerminimierung

- **Beispiele für Maximum-Likelihood Schätzung**
- **Maximum-Likelihood und Fehlerminimierung**
- **Kostenfunktionen**

Maximum-Likelihood Schätzung für Regression:

Likelihood für Input-Output-Paare $(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}),\ m = 1,\ldots,M$

$$p(D \mid \mathbf{w}) = \prod_{m=1}^{M} p(\mathbf{y}^{(m)}, \mathbf{x}^{(m)} \mid \mathbf{w}) = \prod_{m=1}^{M} p(\mathbf{y}^{(m)} \mid \mathbf{x}^{(m)}, \mathbf{w})\, p(\mathbf{x}^{(m)}) \quad \Rightarrow L(\mathbf{w}) = -\sum_{m=1}^{M} \ln p(\mathbf{y}^{(m)} \mid \mathbf{x}^{(m)}, \mathbf{w}) + c$$

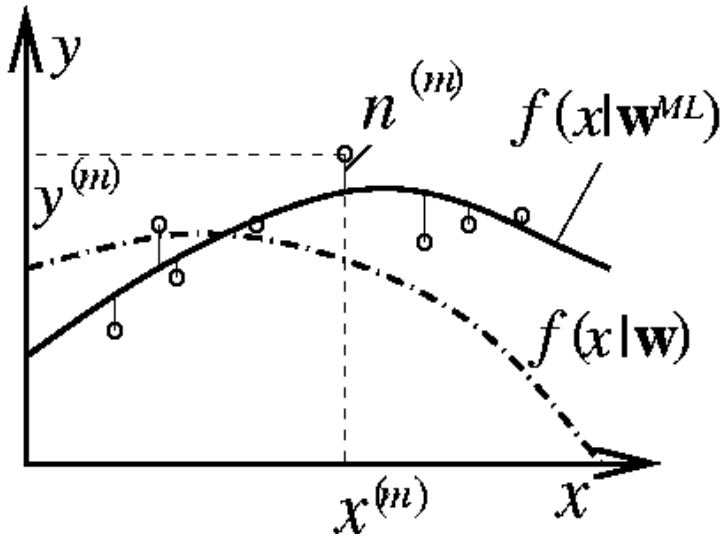
Datenmodell Regression: Alles, was nicht durch f erklärt werden kann, muss Rauschen sein

$$\mathbf{y}^{(m)} = \mathbf{f}(\mathbf{x}^{(m)} \mid \mathbf{w}) + \mathbf{n}^{(m)},\ \Rightarrow \mathbf{n}^{(m)} = \mathbf{y}^{(m)} - \mathbf{f}(\mathbf{x}^{(m)} \mid \mathbf{w})$$

Bei Rauschverteilung $p_n(\mathbf{n})$ **(ohne Konstanten):**

$$L(\mathbf{w}) = -\sum_{m=1}^{M} \ln p(\mathbf{y}^{(m)} \mid \mathbf{x}^{(m)}, \mathbf{w})$$

$$= -\sum_{m=1}^{M} \ln p_n(\mathbf{y}^{(m)} - f(\mathbf{x}^{(m)} \mid \mathbf{w}))$$



Bem: Für Gauss'sches Rauschen gilt:

$$L(\mathbf{w}) = \frac{1}{2\sigma^2} \sum_{m=1}^{M} (\| \mathbf{y}^{(m)} - f(\mathbf{x}^{(m)} \mid \mathbf{w}) \|^2)$$

ML-Schätzung bei Gaussischem Rauschen entspricht dem Least-Squares-Fit

- **Bsp./Wiederholung: Maximum-Likelihood für das lineare Modell**

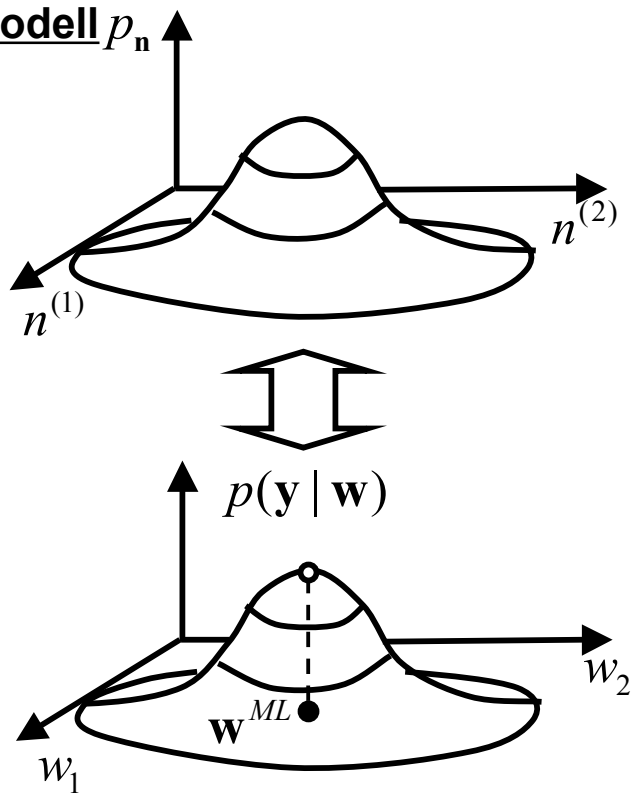
Likelihood: $p(D \mid \mathbf{w}) \equiv p(\mathbf{y} \mid \mathbf{w}) = p_{\mathbf{n}}(\mathbf{y} - \mathbf{H}\mathbf{w})$

Annahme: Gaussisches weißes Rauschen

$$p_{\mathbf{n}}(\mathbf{n}) = \prod_m p_n(n^{(m)}) = \prod_m \frac{1}{(2\pi\sigma_n^2)^{1/2}} \exp\left(-\frac{n^{(m)2}}{2\sigma_n^2}\right)$$

$$= \frac{1}{(2\pi\sigma_n^2)^{M/2}} \exp\left(-\frac{\mathbf{n}^T\mathbf{n}}{2\sigma_n^2}\right)$$

$$p(\mathbf{y} \mid \mathbf{w}) \propto \exp\left(-\frac{(\mathbf{y}-\mathbf{H}\mathbf{w})^T(\mathbf{y}-\mathbf{H}\mathbf{w})}{2\sigma_n^2}\right)$$



- **Maximum-Likelihood Parameter**

$$0! = \nabla_{\mathbf{w}} \ln p(\mathbf{y}|\mathbf{w}) = -\frac{1}{2\sigma_n^2}\nabla_{\mathbf{w}}(\mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T\mathbf{H}^T\mathbf{y} + \mathbf{w}^T\mathbf{H}^T\mathbf{H}\mathbf{w})$$

$$= \frac{1}{\sigma_n^2}(\mathbf{H}^T\mathbf{y} - \mathbf{H}^T\mathbf{H}\mathbf{w})$$

$$= \frac{(\mathbf{H}^T\mathbf{H})}{\sigma_n^2}((\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} - \mathbf{w})$$

$$\Rightarrow \hat{\mathbf{w}} = \mathbf{w}^{ML} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$$

analytisch berechenbar!

- **Beispiel: Hyperparameter im linearen Modell :**

Schätzer Rauschvarianz: $\hat{\sigma}_n^2 = \dfrac{\hat{\mathbf{n}}^T \hat{\mathbf{n}}}{tr(\mathbf{R})}$ $\quad \hat{\mathbf{n}} = \mathbf{y} - \mathbf{H}\hat{\mathbf{w}} = \mathbf{y} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}\mathbf{y} =: \mathbf{R}\mathbf{y}$

- **Bem: Rausch-Schätzung entspricht Schätzung eines Likelihood-Hyperparameters**

Likelihood-Hyperparam.: $p(D \mid \mathbf{w}) = p(D \mid \mathbf{w}, \beta) \quad p(D \mid \beta) = \int p(D \mid \mathbf{w}, \beta)\, p(\mathbf{w} \mid \alpha) d\mathbf{w}$

$$p(\beta \mid D) \equiv \frac{p(D \mid \beta)\, p(\beta)}{\int p(D \mid \beta)\, p(\beta) d\beta}$$

Hier: $p(D \mid \mathbf{w}, \beta) \equiv p(\mathbf{y} \mid \mathbf{w}, \sigma_n^2) = p_n(\mathbf{y} - \mathbf{H}\mathbf{w} \mid \sigma_n^2) \propto \exp\left(-\dfrac{(\mathbf{y} - \mathbf{H}\mathbf{w})^T (\mathbf{y} - \mathbf{H}\mathbf{w})}{2\sigma_n^2} \right)$

MAP/ML: $p(D \mid \beta) \equiv p(\mathbf{y} \mid \sigma_n^2) = \int p(\mathbf{y} \mid \mathbf{w}, \sigma_n^2)\, p(\mathbf{w}) d\mathbf{w} \approx p(\mathbf{y} \mid \mathbf{w}^{ML}, \sigma_n^2)$

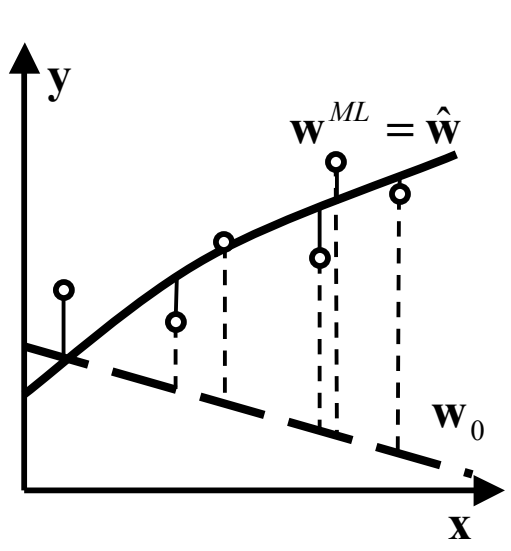
Maximiere bez: σ_n^2 $\quad \dfrac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\dfrac{\hat{\mathbf{n}}^T \hat{\mathbf{n}}}{2\sigma_n^2} \right)$ $\quad$ **(Siehe ML-Bsp):** $\hat{\sigma}_n^2 = \dfrac{\hat{\mathbf{n}}^T \hat{\mathbf{n}}}{M}$

Maximum-Likelihood und Fehlerminimierung

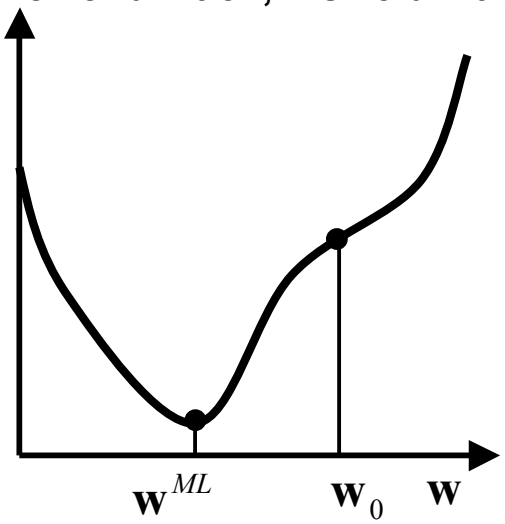
Betrachte überwachtes Lernproblem (Regression, Klassifikation)

- **<u>Likelihood:</u> Gegeben ein Rauschmodell $p_n(\mathbf{n})$, wie wahrscheinlich ist eine Abweichung des tatsächlichen vom vorhergesagten Output?**
- **<u>Negative log-Likelihood:</u> Wie unwahrscheinlich ist die Abweichung**
  Also: Wie groß ist der Fehler, den das Modell macht?

- **Teilweise synonym (beobachtet):**
  ++ negative log-Likelihood
  ++ neg. log. Rauschdichte
  -- „Risikofunktion“
  -- „Fehlerfunktion“
  -- „Verlustfunktion“
  -- „Energiefunktion“
- **Anpassen („Fitten“) eines Modells durch Risiko/Fehlerminimierung**



Fehlerfunktion, Risikofunktion



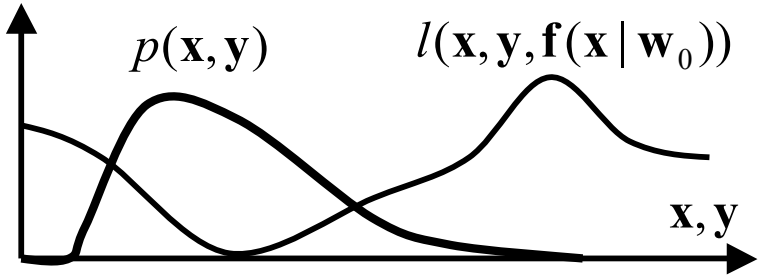
Fehler und empirischer Fehler

- **Def. Fehlerfunktion:** $l(\mathbf{x},\mathbf{y},\mathbf{f}(\mathbf{x}\,|\,\mathbf{w})) \geq 0$ **mit** $l(\mathbf{x},\mathbf{y},\mathbf{y}) = 0 \quad \forall \quad \mathbf{x},\mathbf{y}$

  Bem: -- Es genügt, dass l am wahren Output sein Minimum annimmt

- **Mittlerer (erwarteter) Fehler:**

$$L_e(\mathbf{w}) = \int l(\mathbf{x},\mathbf{y},f(\mathbf{x}\,|\,\mathbf{w}))\,p(\mathbf{x},\mathbf{y})\,d\mathbf{x}\,d\mathbf{y}$$

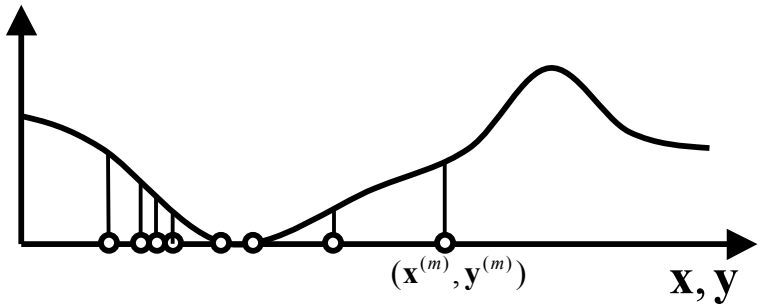


- **Empirischer Fehler:**

  Dichte ist in der Regel unbekannt.

  Man hat Daten $(\mathbf{x}^{(m)},\mathbf{y}^{(m)}),\ m = 1,\ldots,M$

$$L(\mathbf{w}) = \frac{1}{M}\sum_{m=1}^{M} l(\mathbf{x}^{(m)},\mathbf{y}^{(m)},f(\mathbf{x}^{(m)}\,|\,\mathbf{w}))$$



- **Bsp: Quadratischer Fehler**

$$L(\mathbf{w}) = \frac{1}{M}\sum_{m=1}^{M} l(\mathbf{x}^{(m)},\mathbf{y}^{(m)},f(\mathbf{x}^{(m)}\,|\,\mathbf{w})) = \frac{1}{M}\sum_{m=1}^{M}(\mathbf{y}^{(m)} - f(\mathbf{x}^{(m)}\,|\,\mathbf{w}))^2$$

Minimierung des quadratischen Fehlers= „Least Squares Fit“

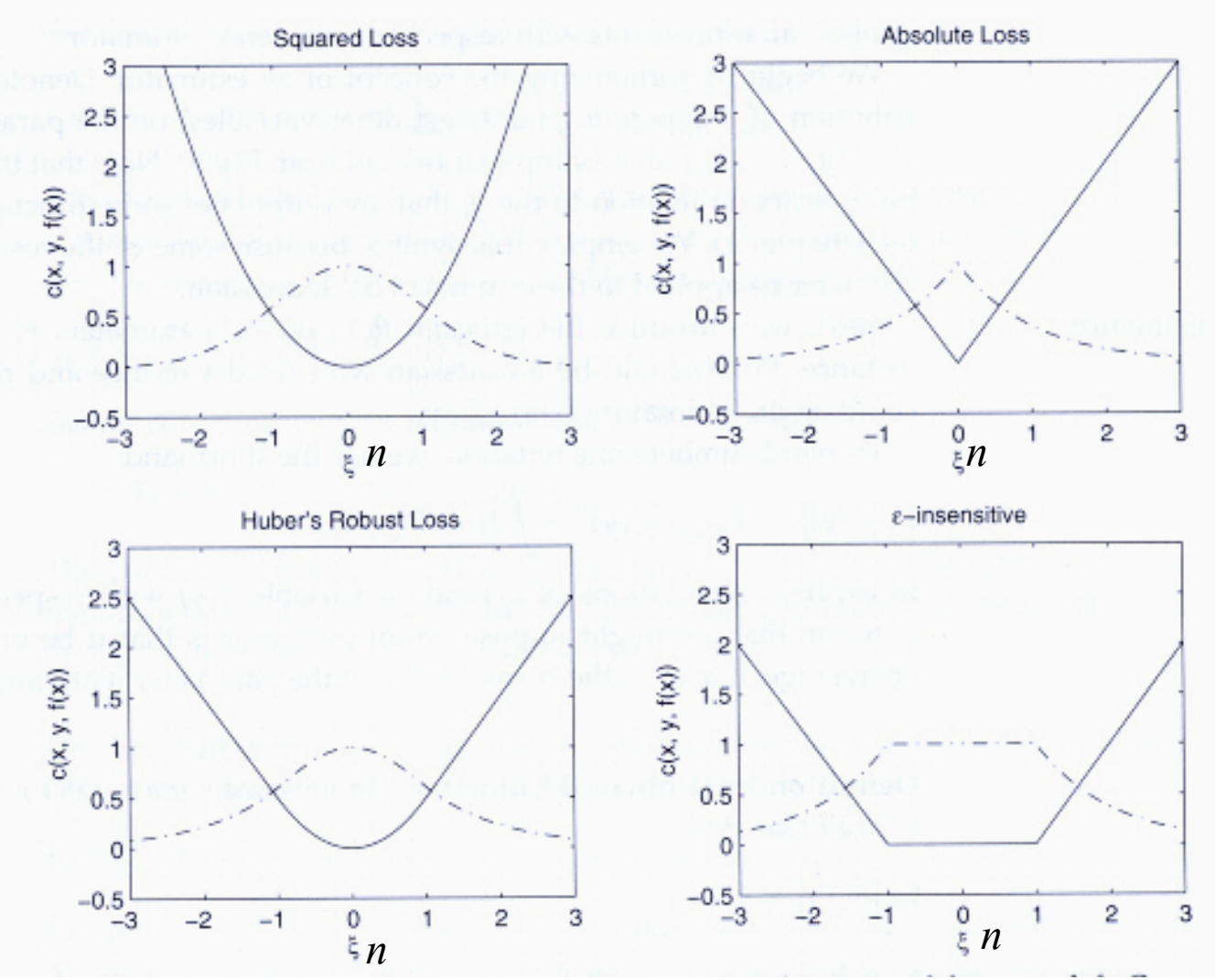
Fehlerfunktionen für Regression

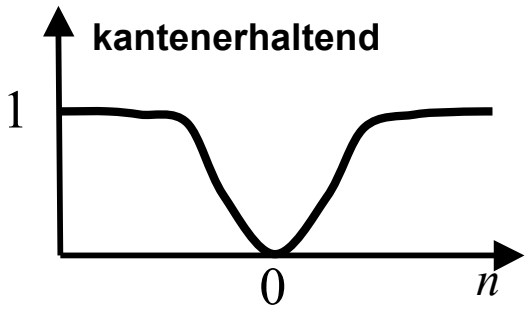
$$l(\mathbf{x},\mathbf{y},\mathbf{f}(\mathbf{x}\,|\,\mathbf{w})) = -\ln p(\mathbf{y}\,|\,\mathbf{x},\mathbf{w}) + c = -\ln p_n(\mathbf{y}-\mathbf{f}(\mathbf{x}\,|\,\mathbf{w})) + c$$

Rauschmodell	Verteilung $p_n(n)$	Fehler	Fehlerfunktion
Gauss	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{n^2}{2})$	**quadratisch**	n^2
Laplace	$\frac{1}{2}\exp(-\lvert n\rvert)$	**absolut (oszilliert)**	$\lvert n\rvert$
Huber's robust	$\propto \begin{cases} \exp(-\frac{n^2}{2\sigma}), & \lvert n\rvert \le \sigma \\ \exp(\frac{\sigma}{2}-\lvert n\rvert) & \text{sonst} \end{cases}$	**Outlier-robust**	$\propto \begin{cases} \frac{1}{2\sigma}n^2 & \lvert n\rvert \le \sigma \\ \lvert n\rvert - \frac{\sigma}{2} & \text{sonst} \end{cases}$
ε-insensitiv	$\propto \begin{cases} 1/2(1+\varepsilon), & \lvert n\rvert \le \varepsilon \\ \exp(\lvert n\rvert-\varepsilon)/2(1+\varepsilon) & \text{sonst} \end{cases}$	**ε-insensitiv**	$\propto \begin{cases} 0, & \lvert n\rvert \le \varepsilon \\ \lvert n\rvert-\varepsilon & \text{sonst} \end{cases}$
	——	**Kantenerhaltend**	$1-\exp(-\frac{n^2}{2\sigma^2})$

Bem: Es gibt Fehlerfunktionen, für die kein äquivalentes Rauschmodell existiert

Fehlerfunktionen: Beispiele für Regression





Fehlerfunktionen für binäre Klassifikation

- **Likelihood: Wie groß ist die Wahrscheinlichkeit einer korrekten Klassenzuteilung?**

$$\Pr(y=1 \mid f(\mathbf{x} \mid \mathbf{w})) \equiv \Pr(y \mid \mathbf{x}, \mathbf{w})$$

- **Fehlerfunktion:**

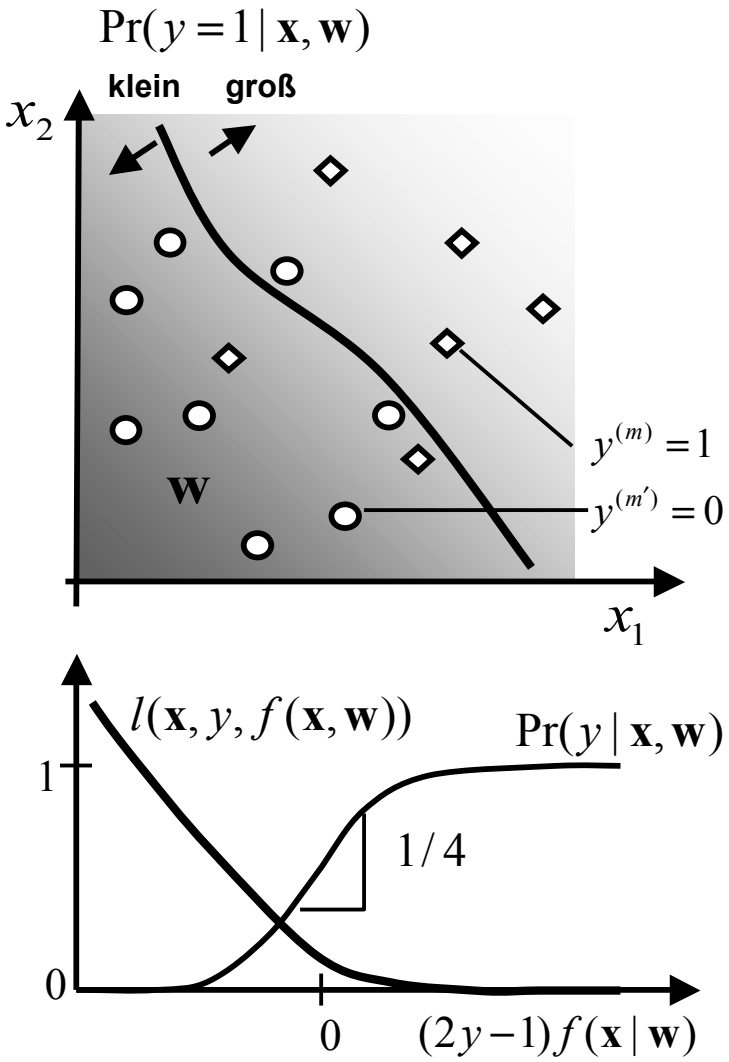
$$l(\mathbf{x}, y, f(\mathbf{x} \mid \mathbf{w})) = -\ln \Pr(y \mid \mathbf{x}, \mathbf{w})$$

- **Logistische Fehlerfunktion:**

$$\Pr(y=1 \mid \mathbf{x}, \mathbf{w}) = \frac{1}{1+\exp(-f(\mathbf{x} \mid \mathbf{w}))}$$

$$l(\mathbf{x}, y=1, f(\mathbf{x} \mid \mathbf{w})) = \ln(1+\exp(-f(\mathbf{x} \mid \mathbf{w}))$$

- **Logistischer empirischer Fehler:**

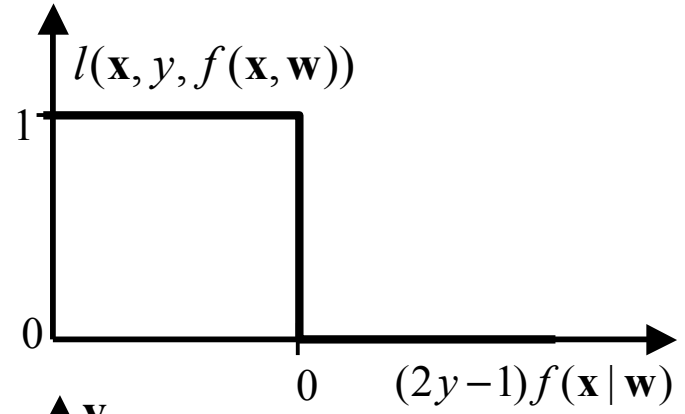


$$L(\mathbf{w}) = \sum_{y^{(m)}=1} \ln(1+\exp(-f(\mathbf{x}^{(m)} \mid \mathbf{w})) + \sum_{y^{(m)}=-1} \ln(1+\exp(+f(\mathbf{x}^{(m)} \mid \mathbf{w})) = \sum_{m=1}^{M} \ln(1+\exp(-(2y^{(m)}-1) f(\mathbf{x}^{(m)} \mid \mathbf{w}))$$

Weitere Fehlerfunktionen

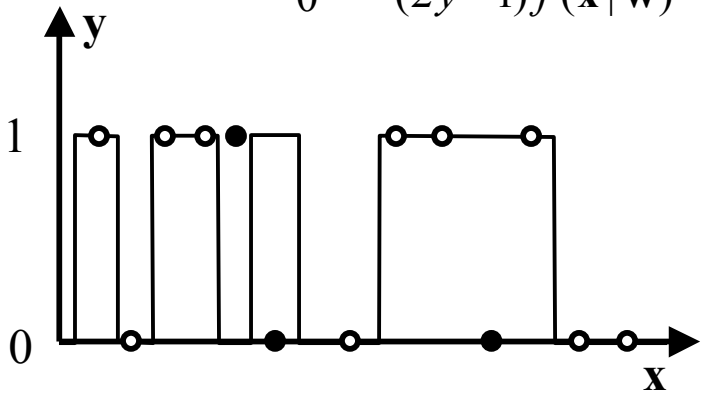
- **Falsch-Klassifikations-Zähler**

$$l(\mathbf{x}, y, f(\mathbf{x}\,|\,\mathbf{w})) = \begin{cases} 0, & (2y-1) f(\mathbf{x}\,|\,\mathbf{w}) \geq 0 \\ 1, & \text{sonst} \end{cases}$$



- **Inputabhängiger Falsch-Klassifikations-Zähler (Bsp: Klassifikation Steine und Diamanten)**

$$l(\mathbf{x}, y, f(\mathbf{x}\,|\,\mathbf{w})) = \begin{cases} 0, & y = f(\mathbf{x}\,|\,\mathbf{w}) \\ l_0(\mathbf{x}), & \text{sonst} \end{cases}$$



- **Outputabhängiger Falsch-Klassifikations-Zähler (Bsp: Klassifikation Krebs/gesund)**

$$l(\mathbf{x}, y, f(\mathbf{x}\,|\,\mathbf{w})) = \begin{cases} 0, y = f(\mathbf{x}\,|\,\mathbf{w}) \\ l_1, y = 1, f(\mathbf{x}\,|\,\mathbf{w}) = 0 \\ l_2, y = 0, f(\mathbf{x}\,|\,\mathbf{w}) = 1 \end{cases}$$

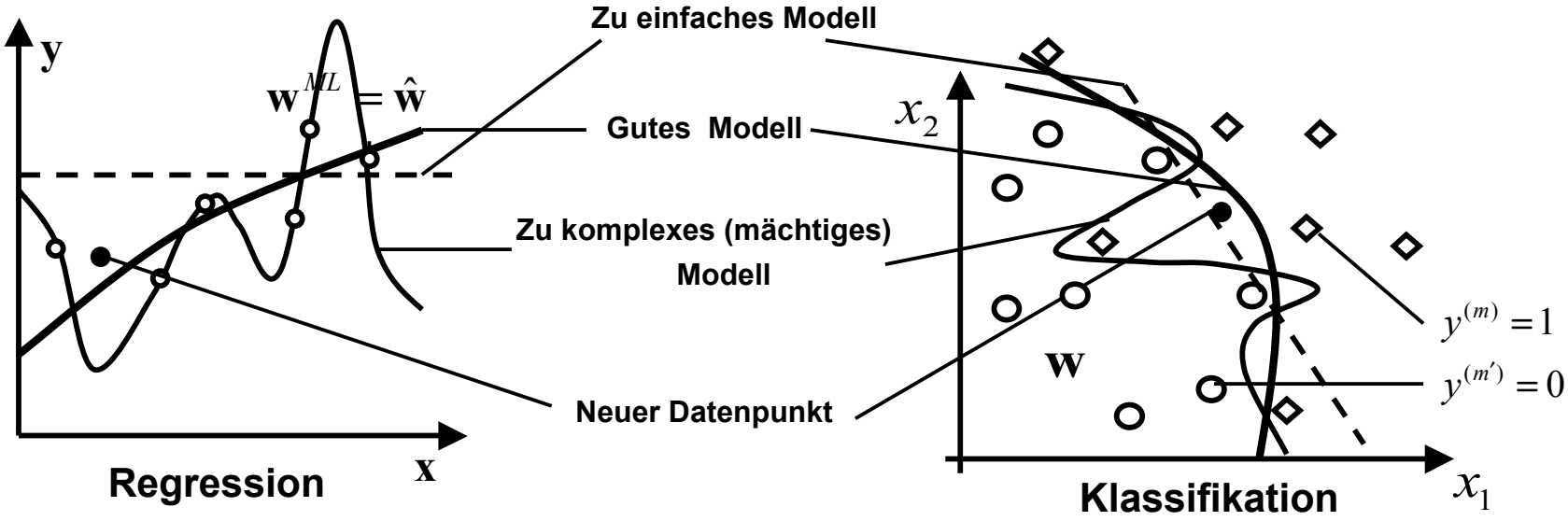
Lernen von Datenmodellen

Generalisierung und Regularisierung

- **Modell-Komplexität und Varianz-Bias Problem**
- **Generalisierung durch Regularisierung**
- **Kreuzvalidierung: Optimierung der Modellkomplexität**
- **Anwendungsbeispiel für Regularisierung: Funktionelle Kernspin-Bilder**

Generalisierung und Regularisierung

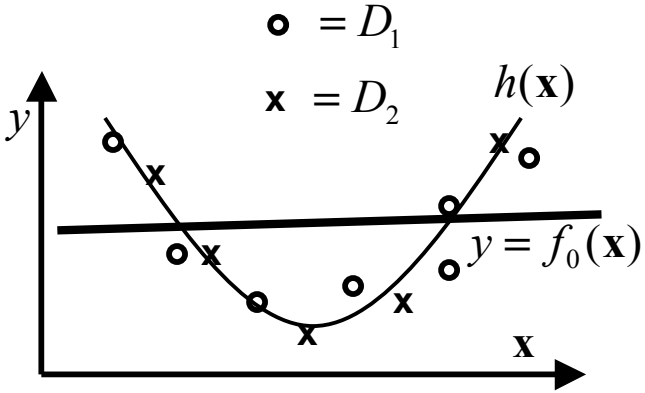
- **Betrachte überwachtes Lernproblem mit endlichem Datensatz** $(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}),\ m = 1, \ldots, M$



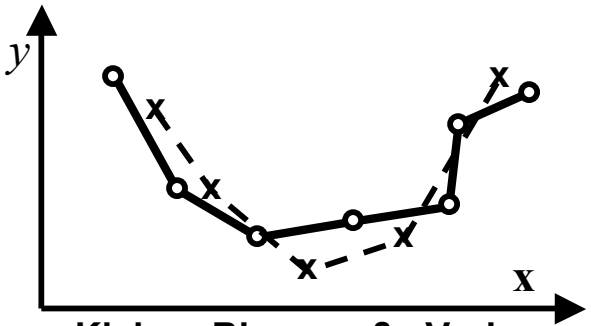
- **Beobachtungen:**
  - **-- Ein genügend komplexes Modell kann die Trainingsdaten beliebig genau erklären (kleiner Trainingsfehler)**
  - **-- Neue Datenpunkte werden dann aber schlechter erklärt (großer Testfehler, schlechte Generalisierungsfähigkeit, „Overfitting“)**
- **Finden der statistischen Struktur: Erkläre Daten mit möglichst einfachem Modell**

Etwas formaler: Das Bias-Varianz Problem

- **Ziel der Datenmodellierung:**
  -- Modellierung der statistischen Struktur h
  -- nicht Modellierung eines speziellen Datensatzes
- **Betrachte mittleres Verhalten über viele Datensätze:**
  $\mathbf{D} = \{D_1, D_2, \ldots\}$
- **Bias: Wie stark weicht das über alle Datensätze gemittelte Modell von dem wahren Modell ab?**
- **Varianz: Wie stark variiert das Modell (wie stark hängt es vom einzelnen Datensatz ab?)**
- **Sei $E_{\mathbf{D}}[.]$ der Erwartungswert über alle Datensätze**
  Für Regression und quadratischen Fehler:



Großer Bias, Varianz=0



Kleiner Bias, große Varianz

$$(f(\mathbf{x}\,|\,\mathbf{w}) - h(\mathbf{x}))^2 = (f(\mathbf{x}\,|\,\mathbf{w}) - E_{\mathbf{D}}[f(\mathbf{x}\,|\,\mathbf{w})] + E_{\mathbf{D}}[f(\mathbf{x}\,|\,\mathbf{w})] - h(\mathbf{x}))^2$$

$$E_{\mathbf{D}}[(f(\mathbf{x}\,|\,\mathbf{w}) - h(\mathbf{x}))^2] = (E_{\mathbf{D}}[f(\mathbf{x}\,|\,\mathbf{w})] - h(\mathbf{x}))^2 + E_{\mathbf{D}}[(f(\mathbf{x}\,|\,\mathbf{w}) - E_{\mathbf{D}}[f(\mathbf{x}\,|\,\mathbf{w})])^2]$$

Minimiere gemeinsam: Bias Varianz

Bias-Varianz-Trade-off in Bayes-Modellen

- **Betrachte Bayes-Schätzer:** $p(\mathbf{w} \mid D) = \frac{1}{p(D)} p(D \mid \mathbf{w}) p(\mathbf{w})$
- **MAP:**
- **Maximierung der Likelihood: Erzielung eines kleinen Bias (guter Fit der Datenpunkte)**
- **Der Prior kann zu komplexe Modelle bestrafen („Occam's Razor")**
  -- Maximierung des Prior favorisiert einfache Modelle
  -- Erzielung einer kleinen Varianz (kein overfitting)
- **Maximierung der Evidenz: Bestimmung der Hyperparameter**
  z. B. Optimierung der relativen Gewichtung von Bias und Varianz-Minimierung
- **Bsp: Weight-Decay Prior** $p(\mathbf{w}) = \frac{1}{Z(\alpha)} \exp(-\frac{\alpha}{2} \|\mathbf{w}\|^2)$
- **Favorisiert kleine Gewichtswerte**
  => favorisiert glattere Kurven (y ändert sich „langsamer" mit $\mathbf{x}$)
- **Negativer Log-Prior entspricht Regularisierungsterm**
  Hyperparameter α gewichtet den Regularisator

$$-\ln p(\mathbf{w}) \equiv \alpha R(\mathbf{w}) = \frac{\alpha}{2} \|\mathbf{w}\|^2 + c$$

Regularisierung: **Favorisierung von Modellen mit bestimmter (z.B. niedriger) Komplexität**

Bayes-Formalismus

-- negative log-Likelihood

-- negativer log-Prior

-- Max. a Posteriori

Schätztheorie

-- Fehlerfunktion L

-- Regularisierungsfunktion R

-- $F(\mathbf{w}) = L(\mathbf{w}) + \alpha R(\mathbf{w}) =! \min$

Beobachtung

-- Rauschen enthält alle (auch hohe) Frequenzen

-- Deterministische Zusammenhänge sind oft glatt (niedrige Frequenzen)

-- Sinnvolle Regularisierung: Bevorzuge „glatte“ Modelle

-- Glattere Modelle <= weniger Parameter, oder kleinere Gewichte

Beispiele für Regularisierungsfunktionen

- **Weight-Decay** $R(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$
- **Kurven-basierte Regularisierung** $R(\mathbf{w}) = \sum_{m=1}^{M} \sum_{j=1}^{c} \sum_{i=1}^{d} \left(\frac{\partial^2 f(\mathbf{x} \mid \mathbf{w})}{\partial x_i \partial x_j} \bigg|_{\mathbf{x}^{(m)}} \right)^2$