Wissensentdeckung in Datenbanken

Optimierung, Überanpassung

Nico Piatkowski und Uwe Ligges

Informatik—Künstliche Intelligenz
Computergestützte Statistik
Technische Universität Dortmund

04.05.2017

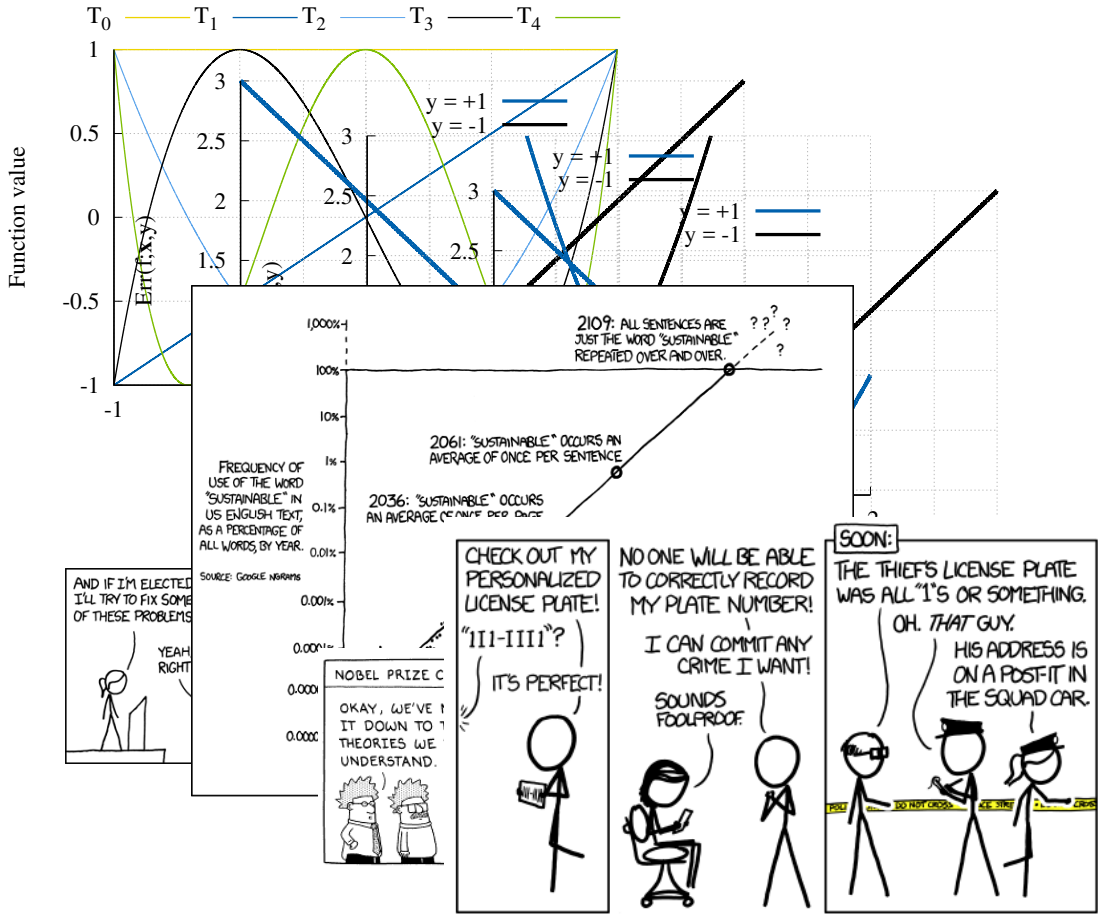
Überblick

Was bisher geschah...

- Modellklassen
- Verlustfunktionen

Heute

- Optimierung
- Overfitting



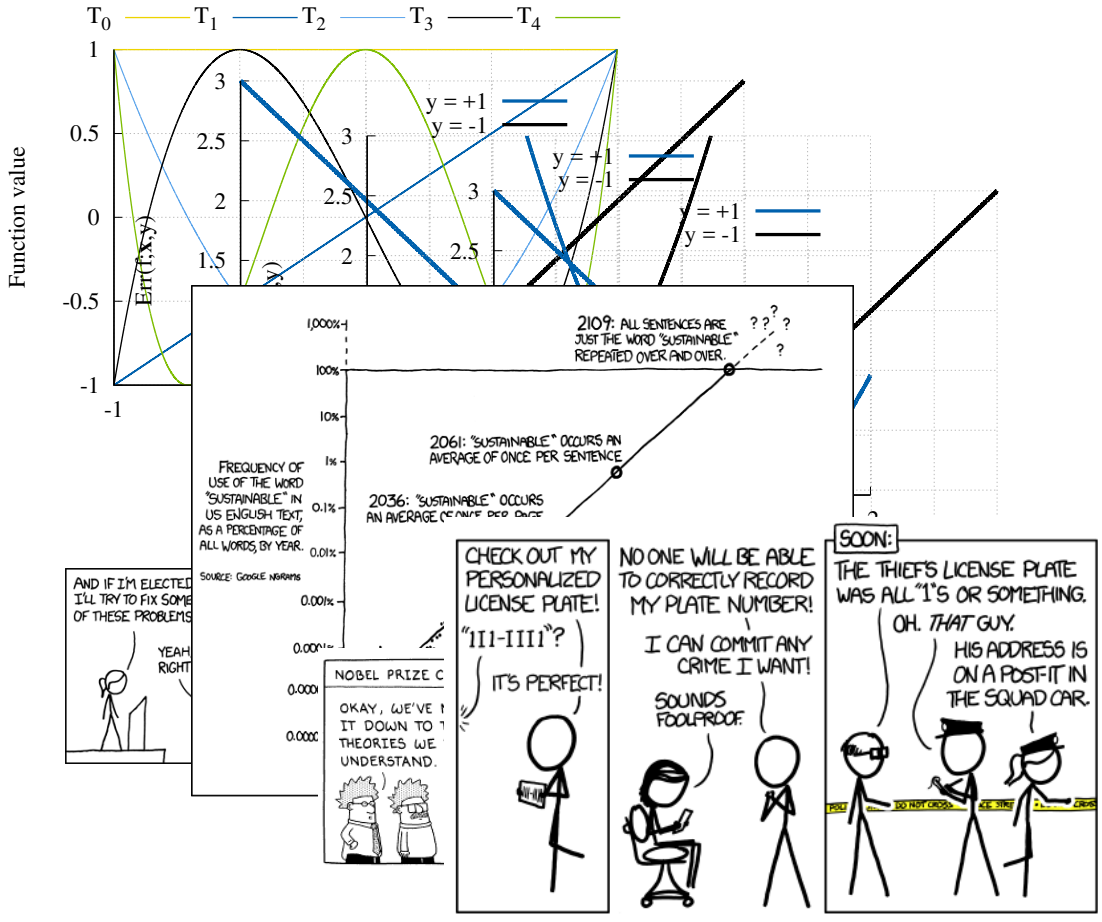
Überblick

Was bisher geschah...

- Modellklassen
- Verlustfunktionen

Heute

- Optimierung
- Overfitting



Differenzierbarkeit

- Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ist *partiell differenzierbar* an β_i, falls

$$\forall \boldsymbol{\beta} \in \mathbb{R}^n : \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_i} = \lim_{h \to 0} \frac{f(\boldsymbol{\beta} + h\boldsymbol{e}_i) - f(\boldsymbol{\beta})}{h}$$

- Der Vektor

$$\nabla f(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_2} \\ \dots \\ \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_n} \end{pmatrix}$$

heißt *Gradient* von f an der Stelle β.

Differenzierbarkeit

- Die Funktion $f : \mathbb{R}^n \to \mathbb{R}$ ist *partiell differenzierbar* an β_i, falls

$$\forall \boldsymbol{\beta} \in \mathbb{R}^n : \frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_i} = \lim_{h \to 0} \frac{f(\boldsymbol{\beta} + h\boldsymbol{e}_i) - f(\boldsymbol{\beta})}{h}$$

- Der Vektor

$$\nabla f(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1} \\ \frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} \\ \cdots \\ \frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_n} \end{pmatrix}$$

heißt *Gradient* von f an der Stelle $\boldsymbol{\beta}$.

Konvexität

- f *konvex*, gdw.

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : f(\boldsymbol{a}) \geq f(\boldsymbol{b}) + \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle$$

- f konvex, gdw.

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \langle \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle \geq 0$$

  $\equiv$ Gradient ist monotoner Operator

- f konvex $\Rightarrow$ Jedes lokale Minimum ist ein globales Minimum

(Neue Literatur im Moodle: Stephen Boyd und Lieven Vandenberghe: Convex Optimization)

Konvexität

- f *konvex*, gdw.

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : f(\boldsymbol{a}) \geq f(\boldsymbol{b}) + \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle$$

- f konvex, gdw.

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \langle \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle \geq 0$$

  $\equiv$ Gradient ist monotoner Operator

- f konvex $\Rightarrow$ Jedes lokale Minimum ist ein globales Minimum

(Neue Literatur im Moodle: Stephen Boyd und Lieven Vandenberghe: Convex Optimization)

Konvexität

- f *konvex*, gdw.

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : f(\boldsymbol{a}) \geq f(\boldsymbol{b}) + \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle$$

- f konvex, gdw.

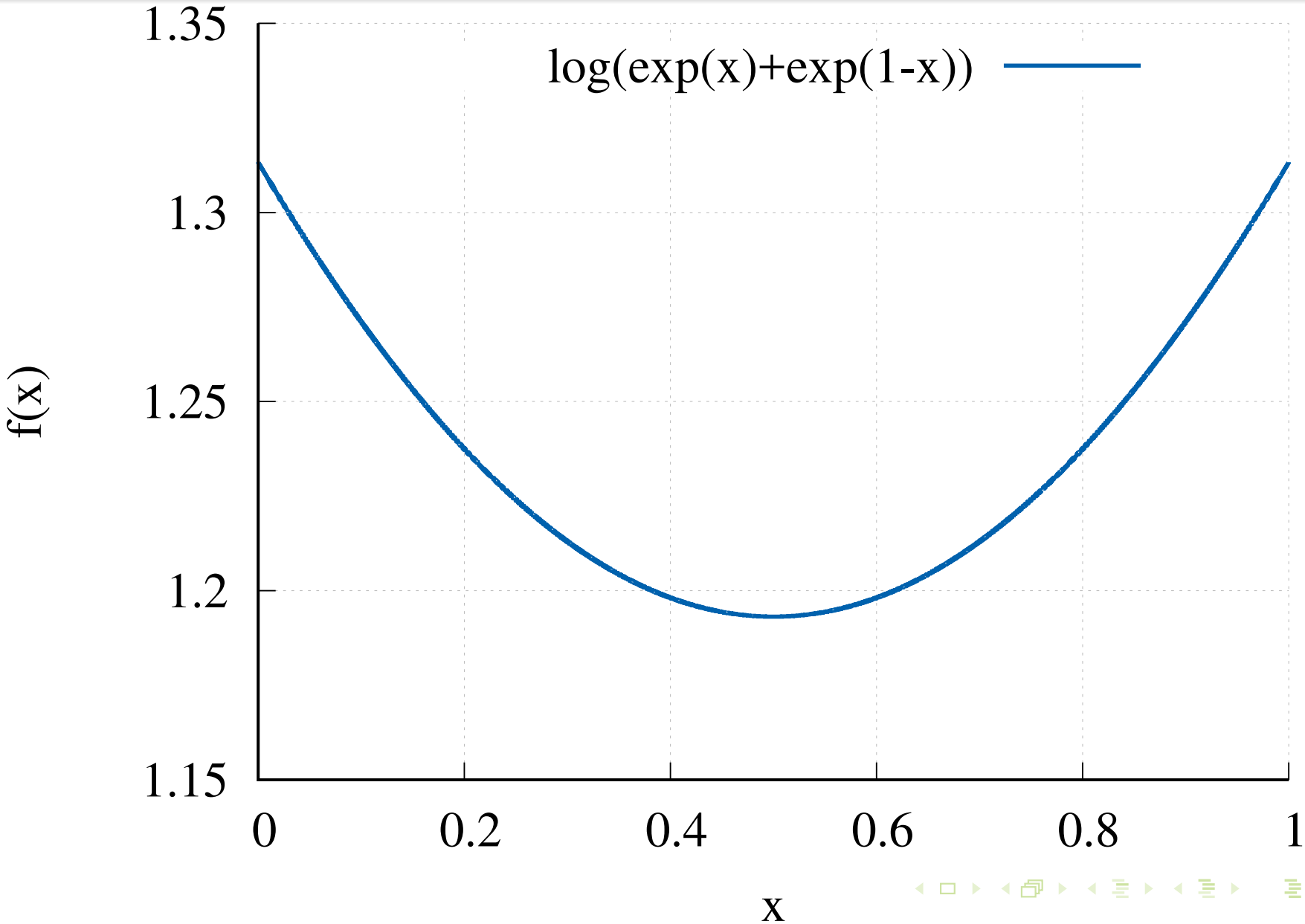
$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \langle \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle \geq 0$$

  $\equiv$ Gradient ist monotoner Operator

- f konvex $\Rightarrow$ Jedes lokale Minimum ist ein globales Minimum

(Neue Literatur im Moodle: Stephen Boyd und Lieven Vandenberghe: Convex Optimization)

Konvexität (II)



Lipschitz-Stetigkeit für $\mathbb{R}^n \to \mathbb{R}^m$

- f heißt *Lipschitz-Stetig mit Konstante* $L > 0$, falls

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : |f(\boldsymbol{a}) - f(\boldsymbol{b})| \leq L \|\boldsymbol{a} - \boldsymbol{b}\|_2$$

Hilfreich bei $f : \mathbb{R}^n \to \mathbb{R}^m$:

- Falls

$$K = \sup_{\boldsymbol{c} \in \mathbb{R}^n} \|\nabla f(\boldsymbol{c})\|_2 < \infty \ ,$$

dann ist f Lipschitz stetig mit Konstante $K > 0$.

Lipschitz-Stetigkeit für $\mathbb{R}^n \to \mathbb{R}^m$

- f heißt *Lipschitz-Stetig mit Konstante* $L > 0$, falls

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : |f(\boldsymbol{a}) - f(\boldsymbol{b})| \leq L \|\boldsymbol{a} - \boldsymbol{b}\|_2$$

Hilfreich bei $f : \mathbb{R}^n \to \mathbb{R}^m$:

- Falls

$$K = \sup_{\boldsymbol{c} \in \mathbb{R}^n} \|\nabla f(\boldsymbol{c})\|_2 < \infty \ ,$$

dann ist f Lipschitz stetig mit Konstante $K > 0$.

Konvexität + Lipschitz stetige Gradienten

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \|\nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b})\|_2 \leq L\|\boldsymbol{a} - \boldsymbol{b}\|_2$$

Multiplikation mit $\|a - b\|_2$ und Cauchy-Schwarz Ungl.:

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \langle \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle \leq L\|\boldsymbol{a} - \boldsymbol{b}\|_2^2$$

Substitution von $g(\boldsymbol{x}) = (L/2)\|\boldsymbol{x}\|_2^2 - f(\boldsymbol{x})$ mit $\nabla g(\boldsymbol{x}) = L\boldsymbol{x} - \nabla f(\boldsymbol{x})$ zeigt g ist konvex. Substitution von $g(\boldsymbol{x})$ in

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : g(\boldsymbol{a}) \geq g(\boldsymbol{b}) + \langle \nabla g(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle$$

führt zu

$$f(\boldsymbol{a}) \leq f(\boldsymbol{b}) + \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + (L/2)\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \qquad (1)$$

Konvexität + Lipschitz stetige Gradienten

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \|\nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b})\|_2 \leq L\|\boldsymbol{a} - \boldsymbol{b}\|_2$$

Multiplikation mit $\|\boldsymbol{a} - \boldsymbol{b}\|_2$ und Cauchy-Schwarz Ungl.:

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \langle \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle \leq L\|\boldsymbol{a} - \boldsymbol{b}\|_2^2$$

Substitution von $g(\boldsymbol{x}) = (L/2)\|\boldsymbol{x}\|_2^2 - f(\boldsymbol{x})$ mit $\nabla g(\boldsymbol{x}) = L\boldsymbol{x} - \nabla f(\boldsymbol{x})$ zeigt g ist konvex. Substitution von $g(\boldsymbol{x})$ in

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : g(\boldsymbol{a}) \geq g(\boldsymbol{b}) + \langle \nabla g(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle$$

führt zu

$$f(\boldsymbol{a}) \leq f(\boldsymbol{b}) + \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + (L/2)\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \qquad (1)$$

Konvexität + Lipschitz stetige Gradienten

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \|\nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b})\|_2 \leq L\|\boldsymbol{a} - \boldsymbol{b}\|_2$$

Multiplikation mit $\|\boldsymbol{a} - \boldsymbol{b}\|_2$ und Cauchy-Schwarz Ungl.:

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \langle \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle \leq L\|\boldsymbol{a} - \boldsymbol{b}\|_2^2$$

Substitution von $g(\boldsymbol{x}) = (L/2)\|\boldsymbol{x}\|_2^2 - f(\boldsymbol{x})$ mit $\nabla g(\boldsymbol{x}) = L\boldsymbol{x} - \nabla f(\boldsymbol{x})$ zeigt g ist konvex. Substitution von $g(\boldsymbol{x})$ in

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : g(\boldsymbol{a}) \geq g(\boldsymbol{b}) + \langle \nabla g(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle$$

führt zu

$$f(\boldsymbol{a}) \leq f(\boldsymbol{b}) + \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + (L/2)\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \qquad (1)$$

Konvexität + Lipschitz stetige Gradienten

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \|\nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b})\|_2 \leq L\|\boldsymbol{a} - \boldsymbol{b}\|_2$$

Multiplikation mit $\|\boldsymbol{a} - \boldsymbol{b}\|_2$ und Cauchy-Schwarz Ungl.:

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \langle \nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle \leq L\|\boldsymbol{a} - \boldsymbol{b}\|_2^2$$

Substitution von $g(\boldsymbol{x}) = (L/2)\|\boldsymbol{x}\|_2^2 - f(\boldsymbol{x})$ mit $\nabla g(\boldsymbol{x}) = L\boldsymbol{x} - \nabla f(\boldsymbol{x})$ zeigt g ist konvex. Substitution von $g(\boldsymbol{x})$ in

$$\forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : g(\boldsymbol{a}) \geq g(\boldsymbol{b}) + \langle \nabla g(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle$$

führt zu

$$f(\boldsymbol{a}) \leq f(\boldsymbol{b}) + \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + (L/2)\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \qquad (1)$$

Methoden

- Spezialisiert für bestimmte Modellklassen
  - Stützvektormethode und Sequential Minimal Optimization
  - Markov Random Fields und Iterative Proportional Fitting
- Generisch
  - Erster Ordnung, mit Gradient
    - Einfache Implementierung
    - Geringer Ressourcenverbrauch
    - Vergleichsweise langsame Konvergenz
  - Zweiter Ordnung, mit Hesse-Matrix
    - Hoher Ressourcenverbrauch
    - Schnelle Konvergenz
  - Proximal-Point Methoden Erster Ordnung
    - Unterstützung für Nebenbedingungen und nicht-überall differenzierbare Funktionen
  - Heuristiken, mit Glück (Evolutionäre Algorithmen, Randomisierte Suche)

Methoden

- Spezialisiert für bestimmte Modellklassen
  - Stützvektormethode und Sequential Minimal Optimization
  - Markov Random Fields und Iterative Proportional Fitting
- Generisch
  - Erster Ordnung, mit Gradient
    - Einfache Implementierung
    - Geringer Ressourcenverbrauch
    - Vergleichsweise langsame Konvergenz
  - Zweiter Ordnung, mit Hesse-Matrix
    - Hoher Ressourcenverbrauch
    - Schnelle Konvergenz
  - Proximal-Point Methoden Erster Ordnung
    - Unterstützung für Nebenbedingungen und nicht-überall differenzierbare Funktionen
  - Heuristiken, mit Glück (Evolutionäre Algorithmen, Randomisierte Suche)

Methoden

- Spezialisiert für bestimmte Modellklassen
  - Stützvektormethode und Sequential Minimal Optimization
  - Markov Random Fields und Iterative Proportional Fitting
- Generisch
  - Erster Ordnung, mit Gradient
    - Einfache Implementierung
    - Geringer Ressourcenverbrauch
    - Vergleichsweise langsame Konvergenz
  - Zweiter Ordnung, mit Hesse-Matrix
    - Hoher Ressourcenverbrauch
    - Schnelle Konvergenz
  - Proximal-Point Methoden Erster Ordnung
    - Unterstützung für Nebenbedingungen und nicht-überall differenzierbare Funktionen
  - Heuristiken, mit Glück (Evolutionäre Algorithmen, Randomisierte Suche)

Methoden

- Spezialisiert für bestimmte Modellklassen
  - Stützvektormethode und Sequential Minimal Optimization
  - Markov Random Fields und Iterative Proportional Fitting
- Generisch
  - Erster Ordnung, mit Gradient
    - Einfache Implementierung
    - Geringer Ressourcenverbrauch
    - Vergleichsweise langsame Konvergenz
  - Zweiter Ordnung, mit Hesse-Matrix
    - Hoher Ressourcenverbrauch
    - Schnelle Konvergenz
  - Proximal-Point Methoden Erster Ordnung
    - Unterstützung für Nebenbedingungen und nicht-überall differenzierbare Funktionen
  - Heuristiken, mit Glück (Evolutionäre Algorithmen, Randomisierte Suche)

Methoden

- Spezialisiert für bestimmte Modellklassen
  - Stützvektormethode und Sequential Minimal Optimization
  - Markov Random Fields und Iterative Proportional Fitting
- Generisch
  - Erster Ordnung, mit Gradient
    - Einfache Implementierung
    - Geringer Ressourcenverbrauch
    - Vergleichsweise langsame Konvergenz
  - Zweiter Ordnung, mit Hesse-Matrix
    - Hoher Ressourcenverbrauch
    - Schnelle Konvergenz
  - Proximal-Point Methoden Erster Ordnung
    - Unterstützung für Nebenbedingungen und nicht-überall differenzierbare Funktionen
  - Heuristiken, mit Glück (Evolutionäre Algorithmen, Randomisierte Suche)

Gradientenabstieg

Jetzt: Modelle repräsentiert durch Parameter vektor $\boldsymbol{\beta} \in \mathbb{R}^n$.
Lernen von $\boldsymbol{\beta}$ mittels *Gradientenabstieg*:

1. Wähle beliebigen Startwert $\boldsymbol{\beta}^0$ sowie $\eta_0, \eta_1, \eta_2, \ldots,$
2. Erzeuge Sequenz von Modellen $\boldsymbol{\beta}^0, \boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \ldots$ mittels

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \eta_t \nabla \ell(\boldsymbol{\beta}^t; \mathcal{D})$$

Sei $\ell(\boldsymbol{\beta}; \mathcal{D})$ eine konvexe Funktion (in $\boldsymbol{\beta}$), mit Lipschitz stetigem Gradienten (Konstante L) und $\eta_t = \eta = 1/L$. Dann gilt nach t Schritten: $\ell(\boldsymbol{\beta}^t; \mathcal{D}) - \ell(\boldsymbol{\beta}^*; \mathcal{D}) \leq \frac{L}{2t} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\|_2^2$.

Gradientenabstieg

Jetzt: Modelle repräsentiert durch Parameter vektor $\boldsymbol{\beta} \in \mathbb{R}^n$.
Lernen von $\boldsymbol{\beta}$ mittels *Gradientenabstieg*:

1. Wähle beliebigen Startwert $\boldsymbol{\beta}^0$ sowie $\eta_0, \eta_1, \eta_2, \ldots,$
2. Erzeuge Sequenz von Modellen $\boldsymbol{\beta}^0, \boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \ldots$ mittels

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \eta_t \nabla \ell(\boldsymbol{\beta}^t; \mathcal{D})$$

Sei $\ell(\boldsymbol{\beta}; \mathcal{D})$ eine konvexe Funktion (in $\boldsymbol{\beta}$), mit Lipschitz stetigem Gradienten (Konstante L) und $\eta_t = \eta = 1/L$. Dann gilt nach t Schritten: $\ell(\boldsymbol{\beta}^t; \mathcal{D}) - \ell(\boldsymbol{\beta}^*; \mathcal{D}) \leq \frac{L}{2t} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\|_2^2$.

Gradientenabstieg—Konvergenz

Sei $\ell(\boldsymbol{\beta}; \mathcal{D})$ eine konvexe Funktion (in $\boldsymbol{\beta}$), mit Lipschitz stetigem Gradienten (Konstante L) und $\eta_t = \eta = 1/L$. Dann gilt nach t Schritten,

$$\ell(\boldsymbol{\beta}^t; \mathcal{D}) - \ell(\boldsymbol{\beta}^*; \mathcal{D}) \leq \frac{L}{2t} \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\|_2^2 \ .$$

Also erfordert $\ell(\boldsymbol{\beta}^t; \mathcal{D}) - \ell(\boldsymbol{\beta}^*; \mathcal{D}) \leq \epsilon$ höchstens $\mathcal{O}(L/\epsilon)$ Schritte.

Gradientenabstieg—Konvergenz (II)

Beweis. Substitution von $\boldsymbol{a} = \boldsymbol{\beta}^{t+1}$ und $\boldsymbol{b} = \boldsymbol{\beta}^{t}$ in (1),

$$\begin{aligned}
\ell(\boldsymbol{\beta}^{t+1}; \mathcal{D}) &\leq \ell(\boldsymbol{\beta}^{t}; \mathcal{D}) - (\eta - (L\eta^2)/2) \|\nabla \ell(\boldsymbol{\beta}^{t}; \mathcal{D})\|_2^2 \\
&\leq \ell(\boldsymbol{\beta}^{*}; \mathcal{D}) + \left\langle \nabla \ell(\boldsymbol{\beta}^{t}; \mathcal{D}), \boldsymbol{\beta}^{t} - \boldsymbol{\beta}^{*} \right\rangle - (\eta/2) \|\nabla \ell(\boldsymbol{\beta}^{t}; \mathcal{D})\|_2^2 \\
&= \ell(\boldsymbol{\beta}^{*}; \mathcal{D}) + (1/(2\eta))(\|\boldsymbol{\beta}^{t} - \boldsymbol{\beta}^{*}\|_2^2 - \|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^{*}\|_2^2) \quad (2)
\end{aligned}$$

Die Sequenz der Funktionswerte $\ell(\boldsymbol{\beta}^{t}; \mathcal{D})$ ist monoton fallend (erste Umformung).

Damit ist $\ell(\boldsymbol{\beta}^{t}; \mathcal{D}) - \ell(\boldsymbol{\beta}^{*}; \mathcal{D})$ kleiner als die übrigen Folgeglieder $\ell(\boldsymbol{\beta}^{j}; \mathcal{D}) - \ell(\boldsymbol{\beta}^{*}; \mathcal{D})$ mit $j < t$.

Gradientenabstieg—Konvergenz (II)

Beweis. Substitution von $\boldsymbol{a} = \boldsymbol{\beta}^{t+1}$ und $\boldsymbol{b} = \boldsymbol{\beta}^{t}$ in (1),

$$\begin{aligned}
\ell(\boldsymbol{\beta}^{t+1}; \mathcal{D}) &\leq \ell(\boldsymbol{\beta}^{t}; \mathcal{D}) - (\eta - (L\eta^2)/2) \|\nabla \ell(\boldsymbol{\beta}^{t}; \mathcal{D})\|_2^2 \\
&\leq \ell(\boldsymbol{\beta}^{*}; \mathcal{D}) + \left\langle \nabla \ell(\boldsymbol{\beta}^{t}; \mathcal{D}), \boldsymbol{\beta}^{t} - \boldsymbol{\beta}^{*} \right\rangle - (\eta/2) \|\nabla \ell(\boldsymbol{\beta}^{t}; \mathcal{D})\|_2^2 \\
&= \ell(\boldsymbol{\beta}^{*}; \mathcal{D}) + (1/(2\eta))(\|\boldsymbol{\beta}^{t} - \boldsymbol{\beta}^{*}\|_2^2 - \|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^{*}\|_2^2) \quad (2)
\end{aligned}$$

Die Sequenz der Funktionswerte $\ell(\boldsymbol{\beta}^{t}; \mathcal{D})$ ist monoton fallend (erste Umformung).

Damit ist $\ell(\boldsymbol{\beta}^{t}; \mathcal{D}) - \ell(\boldsymbol{\beta}^{*}; \mathcal{D})$ kleiner als die übrigen Folgeglieder $\ell(\boldsymbol{\beta}^{j}; \mathcal{D}) - \ell(\boldsymbol{\beta}^{*}; \mathcal{D})$ mit $j < t$.

Gradientenabstieg—Konvergenz (III)

Das Minimum $\ell(\boldsymbol{\beta}^t; \mathcal{D}) - \ell(\boldsymbol{\beta}^*; \mathcal{D})$ ist kleiner als der Mittelwert der vorherigen Folgeglieder $\ell(\boldsymbol{\beta}^j; \mathcal{D}) - \ell(\boldsymbol{\beta}^*; \mathcal{D})$. Darum

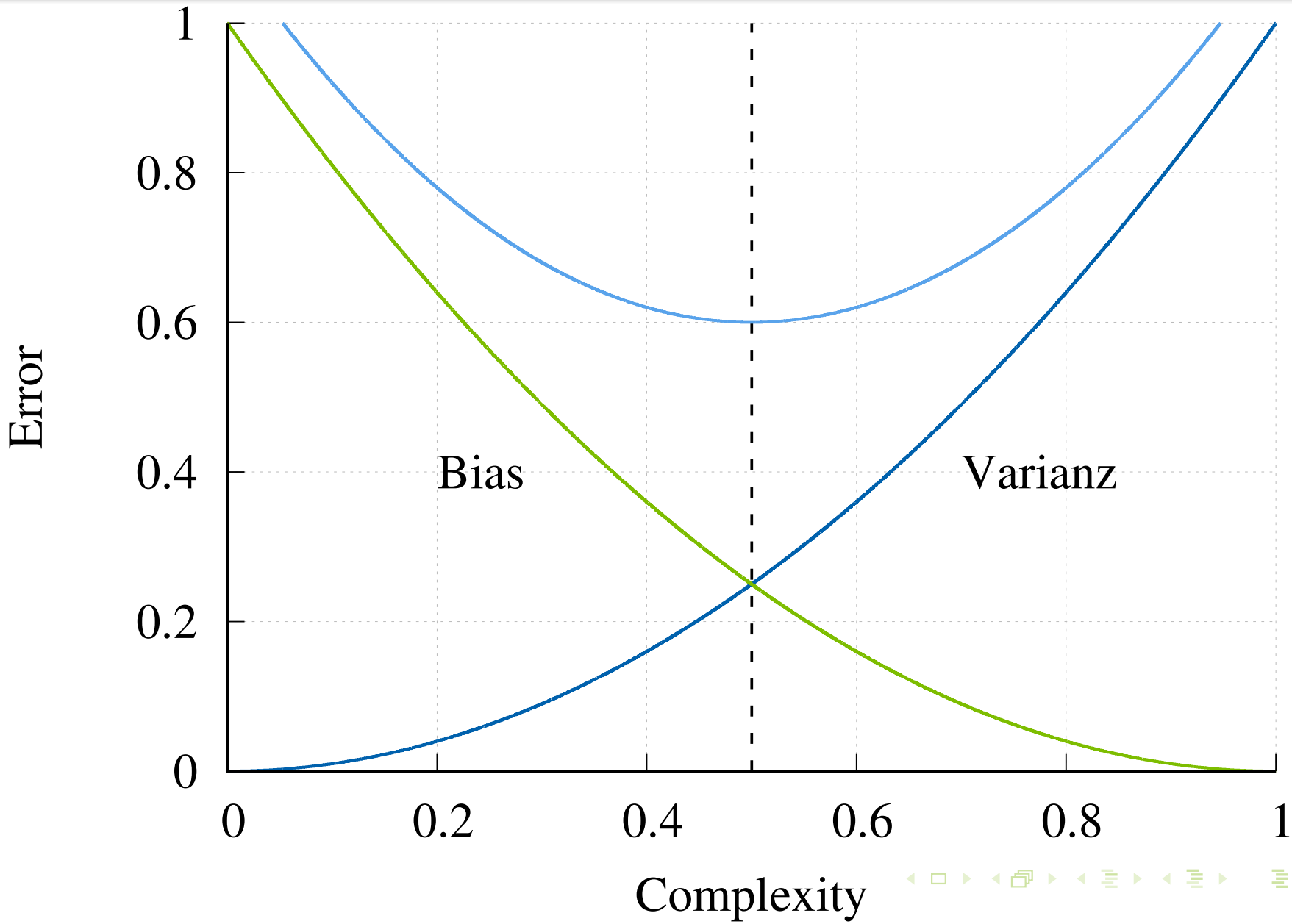
$$\ell(\boldsymbol{\beta}^t; \mathcal{D}) - \ell(\boldsymbol{\beta}^*; \mathcal{D}) \leq \frac{1}{t} \sum_{j=1}^{t-1} \ell(\boldsymbol{\beta}^j; \mathcal{D}) - \ell(\boldsymbol{\beta}^*; \mathcal{D}) \leq \frac{L \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\|_2^2}{2t} \, .$$

Die letzte Ungleichung folgt aus $\eta_t = \eta = 1/L$ und (2). ■

Überanpassung

Überanpassung

Bias und Varianz



Regularisierung

Intuition:

- Modell passt sich ggf. an das Rauschen ϵ in den Daten an
- Optimum der Verlustfunktion $\ell(f; \mathcal{D})$ liefert möglicher suboptimale Vorhersagen

Idee: bestrafe Überanpassung mittels *Regularisierung*
$R : \mathcal{M} \to \mathbb{R}$

$$f^* = \arg\min_{f \in \mathcal{M}} \ell(f; \mathcal{D}) + \lambda R(f)$$

Der Parameter $\lambda > 0$ bestimmt den Einfluss der Regularisierung.

Oft: $R(f) = \|f\|_q$