

# I. Unrestringierte Probleme

## Einführung

### Lösbarkeit

**Definition 1.2.3** (Lösbarkeit).

Das Minimierungsproblem  $P$  heißt **lösbar**, falls ein  $\bar{x} \in M$  existiert mit

$$\inf_{x \in M} f(x) = f(\bar{x})$$

**Satz 1.2.5.** Das Minimierungsproblem  $P$  ist genau dann lösbar, wenn es einen globalen Minimalpunkt besitzt.

**Bemerkung.** Es können drei Fälle der Unlösbarkeit auftreten:

- $\inf_{x \in M} f(x) = +\infty$
- $\inf_{x \in M} f(x) = -\infty$
- Ein endliches Infimum wird nicht angenommen.

**Satz 1.2.6** (Satz von Weierstraß).

Die Menge  $M \subseteq \mathbb{R}^n$  sei nichtleer und kompakt, und die Funktion  $f: M \rightarrow \mathbb{R}$  sei stetig. Dann besitzt  $f$  auf  $M$  (mindestens) einen globalen Minimalpunkt und einen globalen Maximalpunkt.

**Definition 1.2.8** (Untere Niveaumenge). Für  $X \subseteq \mathbb{R}^n$ ,  $f: X \rightarrow \mathbb{R}$  und  $\alpha \in \mathbb{R}$  heißt

$$\text{lev}_{\leq}^{\alpha}(f, X) = \{x \in X \mid f(x) \leq \alpha\}$$

**untere Niveaumenge von  $f$  auf  $X$  zum Niveau  $\alpha$ .** Im Fall  $X = \mathbb{R}^n$  schreiben wir auch kurz

$$f_{\leq}^{\alpha} := \text{lev}_{\leq}^{\alpha}(f, \mathbb{R}^n) = \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$$

**Übung 1.2.10.** Für eine abgeschlossene Menge  $X \subseteq \mathbb{R}^n$  sei die Funktion  $f: X \rightarrow \mathbb{R}$ . Dann ist die Menge  $\text{lev}_{\leq}^{\alpha}(f, X)$  für alle  $\alpha \in \mathbb{R}$  abgeschlossen.

**Übung 1.2.11.** Für eine abgeschlossene Menge  $X \subseteq \mathbb{R}^n$  und endliche Indexmengen  $I$  und  $J$  seien die Funktion  $g_i: X \rightarrow \mathbb{R}, i \in I$ , und  $h_j: X \rightarrow \mathbb{R}, j \in J$ , stetig. Dann ist die Menge

$$M = \{x \in X \mid g_i(x) \leq 0, i \in I, h_j(x) = 0, j \in J\}$$

abgeschlossen.

**Definition.** Die Menge der globalen Minimalpunkte lautet:

$$S = \{\bar{x} \in M \mid \forall x \in M : f(x) \geq f(\bar{x})\}$$

**Lemma 1.2.12.** Für ein  $\alpha \in \mathbb{R}$  sei  $\text{lev}_{\leq}^{\alpha}(f, M) \neq \emptyset$ . Dann gilt

$$S \subseteq \text{lev}_{\leq}^{\alpha}(f, M).$$

**Satz 1.2.13** (Verschärfter Satz von Weierstraß). Für eine (nicht notwendigerweise beschränkte oder abgeschlossene) Menge  $M \subseteq \mathbb{R}^n$  sei  $f: M \rightarrow \mathbb{R}$  stetig, und mit einem  $\alpha \in \mathbb{R}$  sei  $\text{lev}_{\leq}^{\alpha}(f, M)$  nichtleer und kompakt. Dann besitzt  $f$  auf  $M$  (mindestens) einen globalen Minimalpunkt.

**Definition 1.2.21** (Koerzitivität). Gegeben seien eine abgeschlossene Menge  $X \subseteq \mathbb{R}^n$  und eine Funktion  $f: \mathbb{R} \rightarrow \mathbb{R}$  fall für alle Folgen  $(x^k) \subseteq X$  mit  $\lim_k \|x^k\| = +\infty$  auch

$$\lim_k f(x^k) = +\infty$$

gilt, dann heißt  $f$  **koerziv** auf  $X$ .

**Übung 1.2.24.** Gegeben sei die quadratische Funktion  $q(x) = \frac{1}{2}x^T A x + b^T x$  mit einer symmetrischen  $(n, n)$ -Matrix  $A$  (d.h. es gilt  $A = A^T$ ) und  $b \in \mathbb{R}^n$ . Die Funktion  $q$  ist genau dann koerziv auf  $\mathbb{R}^n$ , wenn  $A$  positiv definit ist (d.h. wenn  $d^T A d > 0$  für alle  $d \in \mathbb{R}^n \setminus \{0\}$  gilt).

**Beispiel 1.2.25.** Auf kompakten Mengen  $X$  ist jede Funktion  $f$  trivialerweise koerziv.

**Lemma 1.2.26.** Die Funktion  $f: X \rightarrow \mathbb{R}$  sei stetig und koerziv auf der (nicht notwendigerweise beschränkten) abgeschlossenen Menge  $X \subseteq \mathbb{R}^n$ . Dann ist die Menge  $\text{lev}_{\leq}^{\alpha}(f, X)$  für jedes Niveau  $\alpha \in \mathbb{R}$  kompakt.

**Korollar 1.2.27.** Es sei  $M$  nichtleer und abgeschlossen, aber nicht notwendigerweise beschränkt. Ferner sei die Funktion  $f: M \rightarrow \mathbb{R}$  stetig und koerziv auf  $M$ . Dann besitzt  $f$  auf  $M$  (mindestens) einen globalen Minimalpunkt.

## Rechenregeln und Umformungen

**Übung 1.2.28** (Skalare Vielfache und Summen). Gegeben seien  $M \subseteq \mathbb{R}^n$  und  $f, g: M \rightarrow \mathbb{R}$ . Dann gilt

$$\text{a) } \forall \alpha \geq 0, \beta \in \mathbb{R}: \min_{x \in M} (\alpha f(x) + \beta) = \alpha (\min_{x \in M} f(x)) + \beta$$

$$\text{b) } \forall \alpha < 0, \beta \in \mathbb{R}: \min_{x \in M} (\alpha f(x) + \beta) = \alpha (\max_{x \in M} f(x)) + \beta$$

$$c) \min_{x \in M} (f(x) + g(x)) \geq \min_{x \in M} f(x) + \min_{x \in M} g(x)$$

**Übung 1.2.29** (Separable Zielfunktion auf kartesischem Produkt). Es seien  $X \subseteq \mathbb{R}^n$ ,  $Y \subseteq \mathbb{R}^m$ ,  $f: X \rightarrow \mathbb{R}$  und  $g: Y \rightarrow \mathbb{R}$ . Dann gilt

$$\min_{(x,y) \in X \times Y} (f(x) + g(y)) = \min_{x \in X} f(x) + \min_{y \in Y} g(y)$$

**Übung 1.2.30** (Vertauschung von Minima und Maxima). Es seien  $X \subseteq \mathbb{R}^n$ ,  $Y \subseteq \mathbb{R}^m$ ,  $M = X \times Y$  und  $f: M \rightarrow \mathbb{R}$  gegeben. Dann gilt:

- a)  $\min_{(x,y) \in M} f(x, y) = \min_{x \in X} \min_{y \in Y} f(x, y) = \min_{y \in Y} \min_{x \in X} f(x, y)$
- b)  $\max_{(x,y) \in M} f(x, y) = \max_{x \in X} \max_{y \in Y} f(x, y) = \max_{y \in Y} \max_{x \in X} f(x, y)$
- c)  $\min_{x \in X} \max_{y \in Y} f(x, y) \geq \max_{y \in Y} \min_{x \in X} f(x, y)$

**Übung 1.2.31** (Monotone Transformation). Zu  $M \subseteq \mathbb{R}^n$  und einer Funktion  $f: M \rightarrow Y$  mit  $Y \subseteq \mathbb{R}$  sei  $\psi: Y \rightarrow \mathbb{R}$  eine streng monoton wachsende Funktion. Dann gilt

$$\min_{x \in M} \psi(f(x)) = \psi\left(\min_{x \in M} f(x)\right),$$

und die lokalen bzw. globalen Minimalpunkte stimmen überein.

**Übung 1.2.32** (Epigraphumformulierung). Gegeben seien  $M \subseteq \mathbb{R}^n$  und eine Funktion  $f: M \rightarrow \mathbb{R}$ . Dann sind die Probleme

$$P: \min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } x \in M \quad \text{und} \quad P_{epi}: \min_{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R}} \alpha \text{ s.t. } f(x) \leq \alpha, x \in M$$

äquivalent, d.h. die Minimalwerte stimmen überein und Minimalpunkte entsprechen sich.

**Definition 1.2.33** (Parallelprojektion). Es sei  $M \subseteq \mathbb{R}^n \times \mathbb{R}^m$ . Dann heißt

$$\text{pr}_x M = \{x \in \mathbb{R}^n \mid \exists y \in \mathbb{R}^m : (x, y) \in M\}$$

**Parallelprojektion** von  $M$  (den „ $x$ -Raum“)  $\mathbb{R}^n$ .

**Übung 1.2.34** (Projektionsumformulierung). Gegeben seien  $M \subseteq \mathbb{R}^n \times \mathbb{R}^m$  und eine Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , die nicht von den Variablen aus  $\mathbb{R}^m$  abhängt. Dann sind die Probleme

$$P: \min_{(x,y) \in \mathbb{R}^n \times \mathbb{R}^m} f(x) \text{ s.t. } (x, y) \in M \quad \text{und} \quad P_{proj}: \min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } x \in \text{pr}_x M$$

äquivalent, d.h. die Minimalwerte stimmen überein und Minimalpunkte entsprechen sich.

## Optimalitätsbedingungen

### Abstiegsrichtung

**Definition 2.1.35.** Es seien  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  und  $\bar{x} \in \mathbb{R}^n$ . Ein Vektor  $d \in \mathbb{R}^n$  heißt **Abstiegsrichtung** für  $f$  in  $\bar{x}$ , falls

$$\exists \hat{t} > 0 \forall t \in (0, \hat{t}): f(\bar{x} + td) < f(\bar{x}).$$

**Übung 2.1.36.** Für  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei  $\bar{x}$  ein lokaler Minimalpunkt, dann kann keine Abstiegsrichtung für  $f$  in  $\bar{x}$  existieren.

**Definition 2.1.37.** Gegeben seien  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , ein Punkt  $\bar{x} \in \mathbb{R}^n$  und ein Richtungsvektor  $d \in \mathbb{R}^n$ . Die Funktion

$$\varphi_d: \mathbb{R}^1 \rightarrow \mathbb{R}, t \mapsto f(\bar{x} + td)$$

heißt **eindimensionale Einschränkung** von  $f$  auf die durch  $\bar{x}$  in Richtung  $d$  verlaufende Gerade.

**Bemerkung.** Es gilt  $\varphi_d(0) = f(\bar{x})$  für jede Richtung  $d \in \mathbb{R}^n$ . Daher ist  $d$  genau dann Abstiegsrichtung für  $f$  in  $\bar{x}$ , wenn

$$\exists \hat{t} > 0 \forall t \in (0, \hat{t}): \varphi_d(t) < \varphi_d(0)$$

### Optimalitätsbedingung erster Ordnung

**Definition 2.1.38.** Eine Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  heißt an  $\bar{x} \in \mathbb{R}^n$  in eine Richtung  $d \in \mathbb{R}^n$  **einseitig richtungsdifferenzierbar**, wenn der Grenzwert

$$f'(\bar{x}, d) := \lim_{t \searrow 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t}$$

existiert. Der Wert  $f'(\bar{x}, d)$  heißt dann **einseitige Richtungsableitung**. Die Funktion  $f$  heißt an  $\bar{x}$  **einseitig richtungsdifferenzierbar**, wenn  $f$  an  $\bar{x}$  in jede Richtung  $d \in \mathbb{R}^n$  einseitig richtungsdifferenzierbar ist, und  $f$  heißt **einseitig richtungsdifferenzierbar**, wenn  $f$  an jedem  $\bar{x} \in \mathbb{R}^n$  einseitig richtungsdifferenzierbar ist.

**Lemma 2.1.39.** Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei an  $\bar{x} \in \mathbb{R}^n$  in Richtung  $d \in \mathbb{R}^n$  einseitig richtungsdifferenzierbar mit  $f'(\bar{x}, d) < 0$ . Dann ist  $d$  Abstiegsrichtung für  $f$  in  $\bar{x}$ .

**Lemma 2.1.40.** Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei an einem lokalen Minimalpunkt  $\bar{x} \in \mathbb{R}^n$  einseitig richtungsdifferenzierbar. Dann gilt  $f'(\bar{x}, d) \geq 0$  für jede Richtung  $d \in \mathbb{R}^n$ .

**Definition 2.1.41** (Abstiegsrichtung erster Ordnung). Für eine am Punkt  $\bar{x} \in \mathbb{R}^n$  in Richtung  $d \in \mathbb{R}^n$  einseitig richtungsdifferenzierbare Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  heißt  $d$  **Abstiegsrichtung erster Ordnung**, falls  $f'(\bar{x}, d) < 0$  gilt.

**Definition 2.1.42** (Stationärer Punkt - unrestringierter Fall). Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei an  $\bar{x} \in \mathbb{R}^n$  einseitig richtungsdifferenzierbar. Dann heißt  $\bar{x}$  **stationärer Punkt** von  $f$ , falls  $f'(\bar{x}, d) \geq 0$  für jede Richtung  $d \in \mathbb{R}^n$  gilt.

**Satz 2.1.43** (Kettenregel). Es seien  $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$  differenzierbar an  $\bar{x} \in \mathbb{R}^n$  und  $f: \mathbb{R}^m \rightarrow \mathbb{R}^k$  differenzierbar an  $g(\bar{x}) \in \mathbb{R}^m$ . Dann ist  $f \circ g: \mathbb{R}^n \rightarrow \mathbb{R}^k$  differenzierbar an  $\bar{x}$  mit

$$D(f \circ g)(\bar{x}) = Df(g(\bar{x})) \cdot Dg(\bar{x}).$$

**Lemma 2.1.44.** Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei am Punkt  $\bar{x} \in \mathbb{R}^n$  differenzierbar, und für die Richtung  $d \in \mathbb{R}^n$  gelte  $\langle \nabla f(\bar{x}), d \rangle < 0$ . Dann ist  $d$  Abstiegsrichtung für  $f$  in  $\bar{x}$ .

**Bemerkung 2.1.45.** Ein Vektor  $d$  ist genau dann eine Abstiegsrichtung erster Ordnung für  $f$  in  $\bar{x}$ , wenn  $d$  einen stumpfen Winkel mit dem Gradienten  $\nabla f(\bar{x})$  bildet. Wobei das Skalarprodukt genau dann negativ, wenn sie einen stumpfen Winkel miteinander bilden, und analog ist das Skalarprodukt genau für einen spitzen Winkel bildende Vektoren positiv.

**Übung 2.1.46.** Gegeben seien  $\bar{x} \in \mathbb{R}^n$ , eine endliche Indexmenge  $K$  und an  $\bar{x}$  differenzierbare Funktionen  $f_k: \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $k \in K$ . Die Funktion  $f(x) := \max_{k \in K} f_k(x)$  ist an  $\bar{x}$  einseitig richtungsdifferenzierbar und dass mit  $K_*(\bar{x}) = \{k \in K \mid f_k(\bar{x}) = f(\bar{x})\}$

$$f'(\bar{x}, d) = \max_{k \in K_*(\bar{x})} \langle \nabla f_k(\bar{x}), d \rangle$$

für jede Richtung  $d \in \mathbb{R}^n$  gilt.

**Satz 2.1.47** (Notwendige Optimalitätsbedingung erster Ordnung – Fermat’sche Regel). Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei differenzierbar an einem lokalen Minimalpunkt  $\bar{x} \in \mathbb{R}^n$ . Dann gilt  $\nabla f(\bar{x}) = 0$ .

**Definition 2.1.48** (Kritischer Punkt). Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei an  $\bar{x} \in \mathbb{R}^n$  differenzierbar. Dann heißt  $\bar{x}$  kritischer Punkt von  $f$ , wenn  $\nabla f(\bar{x}) = 0$  gilt.

**Übung 2.1.49.** Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei differenzierbar an einem Punkt  $\bar{x} \in \mathbb{R}^n$ . Der Punkt  $\bar{x}$  ist genau dann stationärer Punkt von  $f$ , wenn er kritischer Punkt von  $f$  ist.

**Definition 2.1.17** (Sattelpunkt). Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei an  $\bar{x} \in \mathbb{R}^n$  differenzierbar. Dann heißt  $\bar{x}$  **Sattelpunkt** von  $f$ , falls  $\bar{x}$  zwar kritischer Punkt von  $f$ , aber weder lokaler Minimal- noch lokaler Maximalpunkt ist.

## Geometrische Eigenschaften von Gradienten

---

**Algorithmus 2.1:** Konzeptioneller Algorithmus zur unrestringierten nichtlinearen Minimierung mit Informationen erster Ordnung

---

**Input :** Lösbares unrestringiertes differenzierbares Optimierungsproblem  $P$

**Output :** Globaler Minimalpunkt  $x^*$  von  $f$  über  $\mathbb{R}^n$

---

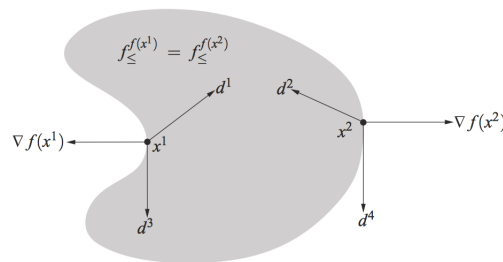
```

1 begin
2   Bestimme alle kritischen Punkte von  $f$ , d. h. die Lösungsmenge  $K$  der Gleichung
    $\nabla f(x) = 0$ .
3   Bestimme einen Minimalpunkt  $x^*$  von  $f$  in  $K$ .
4 end

```

---

**Abb. 2.2** Gradienten und Abstiegsrichtungen



**Bemerkung.** Man kann zeigen, dass jeder Vektor  $d \in \mathbb{R}^n$  mit  $\langle \nabla f(\bar{x}), d \rangle > 0$  eine Anstiegsrichtung erster Ordnung ist. Da für einen nichtkritischen Punkt  $\bar{x}$  die Gradientenrichtung  $d = \nabla f(\bar{x})$  die strikte Ungleichung

$$\langle \nabla f(\bar{x}), \nabla f(\bar{x}) \rangle = \|\nabla f(\bar{x})\|_2^2 > 0$$

erfüllt, ist  $d = \nabla f(\bar{x})$  also eine Anstiegsrichtung erster Ordnung für  $f$  in  $\bar{x}$ , und man kann zeigen, dass  $\nabla f(\bar{x})$  senkrecht auf dem Rand von  $f_{\leq}^{f(\bar{x})}$  steht.

## Optimalitätsbedingungen zweiter Ordnung

**Bemerkung.** Für normierte Richtungen  $d$  liefert die Cauchy-Schwarz-Ungleichung

$$-\|\nabla f(\bar{x})\|_2 = -\|\nabla f(\bar{x})\|_2 \cdot \|d\|_2 \leq \langle \nabla f(\bar{x}), d \rangle \leq \|\nabla f(\bar{x})\|_2 \cdot \|d\|_2 = \|\nabla f(\bar{x})\|_2$$

und die Unter- und Oberschranken werden genau für linear abhängige  $d$  und  $\nabla f(\bar{x})$  angenommen. Wegen  $\nabla f(\bar{x})$  wird die kleinst- und größtmögliche Steigung daher folgend realisiert

$$d_{\min} = -\frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2} \quad \text{und} \quad d_{\max} = \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2}$$

In der Praxis arbeitet man aber nicht mit der negativen Gradientenrichtung, denn gerade in der Nähe der gesuchten kritischen Punkte - nahe bei null - ist die Division  $\frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2}$  numerisch instabil.

**Satz 2.1.19** (Entwicklungen 1. und 2. Ordnung per univariatem Satz von Taylor).

a) Es sei  $\varphi: \mathbb{R} \rightarrow \mathbb{R}$  differenzierbar an  $\bar{t}$ . Dann gilt für alle  $t \in \mathbb{R}$

$$\varphi(t) = \varphi(\bar{t}) + \varphi'(\bar{t})(t - \bar{t}) + o(|t - \bar{t}|),$$

wobei  $o(|t - \bar{t}|)$  einen Ausdruck der Form  $\omega(t) \cdot |t - \bar{t}|$  mit  $\lim_{t \rightarrow \bar{t}} \omega(t) = \omega(\bar{t}) = 0$  bezeichnet.

b) Es sei  $\varphi: \mathbb{R} \rightarrow \mathbb{R}$  zweimal differenzierbar an  $\bar{t}$ . Dann gilt für alle  $t \in \mathbb{R}$

$$\varphi(t) = \varphi(\bar{t}) + \varphi'(\bar{t})(t - \bar{t}) + \frac{1}{2}\varphi''(\bar{t})(t - \bar{t})^2 + o(|t - \bar{t}|^2),$$

wobei  $o(|t - \bar{t}|^2)$  einen Ausdruck der Form  $\omega(t) \cdot |t - \bar{t}|^2$  mit  $\lim_{t \rightarrow \bar{t}} \omega(t) = \omega(\bar{t}) = 0$  bezeichnet.

**Lemma 2.1.20.** Für  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , einen Punkt  $\bar{x} \in \mathbb{R}^n$  und eine Richtung  $d \in \mathbb{R}^n$  seien  $\varphi'_d(0) = 0$  und  $\varphi''_d(0) < 0$ . Dann ist  $d$  Abstiegsrichtung für  $f$  in  $\bar{x}$ .

**Lemma 2.1.21.** Für  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei  $\bar{x}$  ein lokaler Minimalpunkt. Dann gilt  $\nabla f(\bar{x}) = 0$ , und jede Richtung  $d \in \mathbb{R}^n$  erfüllt  $\varphi''_d(0) \geq 0$ .

**Bemerkung.** Es gilt für eine Richtung  $d$  gilt, dass  $\varphi''_d(0) = d^T D^2 f(\bar{x}) d$

**Lemma 2.1.22.** Für  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , einen Punkt  $\bar{x} \in \mathbb{R}^n$  und eine Richtung  $d \in \mathbb{R}^n$  seien  $\langle \nabla f(\bar{x}), d \rangle = 0$  und  $d^T D^2 f(\bar{x}) d < 0$ . Dann ist  $d$  Abstiegsrichtung für  $f$  in  $\bar{x}$ .

**Definition 2.1.23** (Abstiegsrichtung zweiter Ordnung). Zu  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  und  $\bar{x} \in \mathbb{R}^n$  heißt jeder Richtungsvektor  $d \in \mathbb{R}^n$  mit  $\langle \nabla f(\bar{x}), d \rangle = 0$  und  $d^T D^2 f(\bar{x}) d < 0$  **Abstiegsrichtung zweiter Ordnung** für  $f$  in  $\bar{x}$ .

**Satz 2.1.27** (Notwendige Optimalitätsbedingung zweiter Ordnung). Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei zweimal differenzierbar an einem lokalen Minimalpunkt  $\bar{x} \in \mathbb{R}^n$ . Dann gilt  $\nabla f(\bar{x}) = 0$  und  $D^2 f(\bar{x}) \succeq 0$ .

---

**Algorithmus 2.2:** Konzeptioneller Algorithmus zur unrestringierten nichtlinearen Minimierung mit Informationen zweiter Ordnung

---

**Input :** Lösbares unrestringiertes zweimal stetig differenzierbares

Optimierungsproblem  $P$

**Output :** Globaler Minimalpunkt  $x^*$  von  $f$  über  $\mathbb{R}^n$

1 **begin**

2   Bestimme alle kritischen Punkte mit positiv semidefiniter Hesse-Matrix von  $f$ ,  
d. h. die Lösungsmenge  $K$  der beiden Bedingungen  $\nabla f(x) = 0$  und  $D^2 f(x) \succeq 0$ .

3   Bestimme einen Minimalpunkt  $x^*$  von  $f$  in  $K$ .

4 **end**

---

**Satz 2.1.30** (Entwicklungen 1. und 2. Ordnung per univariatem Satz von Taylor).

a) Es sei  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  differenzierbar an  $\bar{x}$ . Dann gilt für alle  $x \in \mathbb{R}^n$

$$f(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + o(\|x - \bar{x}\|),$$

wobei  $o(\|x - \bar{x}\|)$  einen Ausdruck der Form  $\omega(x) \cdot \|x - \bar{x}\|$  mit  $\lim_{x \rightarrow \bar{x}} \omega(x) = \omega(\bar{x}) = 0$  bezeichnet.

b) Es sei  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  zweimal differenzierbar an  $\bar{x}$ . Dann gilt für alle  $x \in \mathbb{R}^n$

$$f(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{1}{2}(x - \bar{x})^T D^2 f(\bar{x})(x - \bar{x}) + o(\|x - \bar{x}\|^2),$$

wobei  $o(\|x - \bar{x}\|^2)$  einen Ausdruck der Form  $\omega(x) \cdot \|x - \bar{x}\|^2$  mit  $\lim_{x \rightarrow \bar{x}} \omega(x) = \omega(\bar{x}) = 0$  bezeichnet.

**Definition.**

- $B_{\leq}(\bar{x}, r) = \{x \in \mathbb{R}^n \mid \|x - \bar{x}\| \leq r\}$
- $B_{=}(\bar{x}, r) = \{x \in \mathbb{R}^n \mid \|x - \bar{x}\| = r\}$

**Satz 2.1.31** (Hinreichende Optimalitätsbedingung zweiter Ordnung). Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei an  $\bar{x} \in \mathbb{R}^n$  zweimal differenzierbar, und es gelte  $\nabla f(\bar{x}) = 0$  und  $D^2 f(\bar{x}) \succ 0$ . Dann ist  $\bar{x}$  ein strikter lokaler Minimalpunkt von  $f$ .

**Definition 2.1.35** (Nichtdegenerierte kritische und Minimalpunkt). Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei an  $\bar{x}$  zweimal differenzierbar mit  $\nabla f(\bar{x}) = 0$ . Dann heißt  $\bar{x}$

- a) **nichtdegenerierter kritischer Punkt**, falls  $D^2 f(\bar{x})$  nichtsingulär ist,
- b) **nichtdegenerierter lokaler Minimalpunkt**, falls  $\bar{x}$  lokaler Minimalpunkt und nichtdegenerierter kritischer Punkt ist.

**Lemma 2.1.36.** Der Punkt  $\bar{x}$  ist genau dann nichtdegenerierter lokaler Minimalpunkt von  $f$ , wenn  $\nabla f(\bar{x}) = 0$  und  $D^2 f(\bar{x}) \succ 0$  gilt.

**Definition.**  $\mathcal{F} = \{f \in C^2(\mathbb{R}^n, \mathbb{R}) \mid \text{alle kritischen Punkte von } f \text{ sind nichtdegeneriert} \}$

**Satz 2.1.37.**  $\mathcal{F}$  ist  $C^2_s$ -offen und -dicht in  $C^2(\mathbb{R}^n, \mathbb{R})$ .

**Übung 2.1.38.** In einem nichtdegeneriertem Sattelpunkt existiert sowohl eine Ab- als auch eine Anstiegsrichtung zweiter Ordnung.

## Konvexe Optimierungsprobleme

**Definition 2.1.39** (Konvexe Mengen und Funktionen).



a) Eine Menge  $X \subseteq \mathbb{R}^n$  heißt **konvex**, falls

$$\forall x, y \in X, \lambda \in (0, 1) : \quad (1 - \lambda)x + \lambda y \in X$$

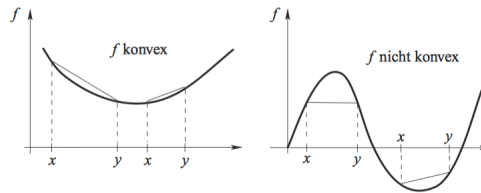
gilt (d.h. die Verbindungsstrecke von je zwei beliebigen Punkten in  $X$  gehört komplett zu  $X$ ).

b) Für eine konvexe Menge  $X \subseteq \mathbb{R}^n$  heißt eine Funktion  $f: X \rightarrow \mathbb{R}$  **konvex** (auf  $X$ ), falls

$$\forall x, y \in X, \lambda \in (0, 1) : \quad f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y)$$

gilt (d.h. der Funktionsgraph von  $f$  verläuft unter jeder seiner Sekanten).

**Bemerkung.** Während die Konvexität einer Funktion geometrisch dadurch definiert ist, dass ihr Graph unter jeder ihrer Sekanten verläuft, lässt sich Konvexität einer stetig differenzierbaren Funktion  $f$  dadurch charakterisieren, dass ihr Graph über den Graphen jeder ihrer Linearisierungen verläuft.



**Abb. 2.4** Konvexität von Funktionen auf  $\mathbb{R}$

**Satz 2.1.40** ( $C^1$ -Charakterisierung von Konvexität). Auf einer konvexen Menge  $X \subseteq \mathbb{R}^n$  ist eine Funktion  $f \in C^1(X, \mathbb{R})$  genau dann konvex, wenn folgendes gilt:

$$\forall x, y \in X : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

**Korollar 2.1.41.** Die Funktion  $f \in C^1(\mathbb{R}^n, \mathbb{R})$  sei konvex. Dann sind die kritischen Punkte von  $f$  genau die globalen Minimalpunkte von  $f$ .

**Satz 2.1.42** ( $C^2$ -Charakterisierung von Konvexität). Eine Funktion  $f \in C^2(\mathbb{R}^n, \mathbb{R})$  ist genau dann konvex, wenn folgendes gilt:

$$\forall x \in \mathbb{R}^n : \quad D^2 f(x) \succeq 0$$

**Übung 2.1.43.** Gegeben sei die quadratische Funktion  $q(x) = \frac{1}{2}x^T A x + b^T x$  mit  $A = A^T \succ 0$  und  $b \in \mathbb{R}^n$ . Die Funktion  $q$  ist eine auf  $\mathbb{R}^n$  (gleichmäßige) konvexe Funktion und ihr eindeutiger Minimalpunkt

$$x^* = -A^{-1}b$$

mit Minimalwert  $q(x^*) = -\frac{1}{2}b^T A^{-1}b$ .

# Numerische Verfahren

## 2.2.1 Abstiegsverfahren

Zunächst betrachten wir Verfahren, die in jedem Iterationsschritt einen Abstieg im Zielfunktionswert erzeugen, für die also

$$\forall k \in \mathbb{N}_0 : f(x^{k+1}) < f(x^k)$$

gilt. Solche Verfahren können nur „unter sehr unglücklichen Umständen“ gegen lokale Maximalpunkte konvergieren und aus geometrischen Überlegungen heraus ist die Konvergenz gegen Sattelpunkte unwahrscheinlich.

Neben der Stetigkeit der Zielfunktion  $f$  werden wir im gesamten Abschn. 2.2 fordern, dass die untere Niveaumenge  $f_{\leq}^{f(x^0)}$  zum Startpunkt  $x^0 \in \mathbb{R}^n$  beschränkt ist.

**Übung.** Als erste algorithmische Idee könnte man versuchen, die Gleichung  $\nabla f(x) = 0$  mit dem aus der Numerik bekannten Newton-Verfahren

$$x^{k+1} = x^k - \left(D^2 f(x^k)\right)^{-1} \nabla f(x^k), \quad k = 0, 1, 2, \dots$$

Vorteil wäre eine hohe Konvergenzgeschwindigkeit, falls  $x^0$  nahe genug an einer Lösung liegt. Nachteilig ist, dass  $x^0$  nicht in der Nähe einer Lösung zu liegen braucht, dass die Hesse-Matrix  $D^2 f(x^k)$  nicht notwendig invertierbar sein muss und dass das Newton-Verfahren auch gegen lokale Maximalpunkte und Sattelpunkte konvergieren kann.

---

**Algorithmus 2.3:** Allgemeines Abstiegsverfahren

---

**Input :**  $C^1$ -Optimierungsproblem  $P$

**Output :** Approximation  $\bar{x}$  eines kritischen Punkts von  $f$  (falls das Verfahren terminiert; [Korollar 2.2.10](#))

```
1 begin
2   Wähle einen Startpunkt  $x^0$ , eine Toleranz  $\varepsilon > 0$  und setze  $k = 0$ .
3   while  $\|\nabla f(x^k)\| > \varepsilon$  do
4     Wähle  $x^{k+1}$  mit  $f(x^{k+1}) < f(x^k)$ .
5     Ersetze  $k$  durch  $k + 1$ .
6   end
7   Setze  $\bar{x} = x^k$ .
8 end
```

---

**Lemma 2.2.3.** Für beschränktes  $f_{\leq}^{f(x^0)}$  bricht die von Algorithmus 2.3 mit  $\varepsilon = 0$  erzeugte Folge  $(x^k)$  entweder nach endlich vielen Schritten mit einem kritischen Punkt ab, oder sie besitzt mindestens einen Häufungspunkt in  $f_{\leq}^{f(x^0)}$ , und die Folge der Funktionswerte  $(f(x^k))$  ist konvergent.

**Definition 2.2.5** (Effiziente Schrittweiten). Es sei  $(d^k)$  eine Folge von Abstiegsrichtungen erster Ordnung, und  $(t^k)$  erfülle

$$\exists c > 0 \quad \forall k \in \mathbb{N} : f(x^k + t^k d^k) - f(x^k) \leq -c \left( \frac{\langle \nabla f(x^k), d^k \rangle}{\|d^k\|_2} \right)^2$$

Dann heißt  $(t^k)$  **effiziente Schrittweitenfolge** für  $(d^k)$ .

**Satz 2.2.6.** Die Menge  $f_{\leq}^{f(x^0)}$  sei beschränkt,  $(d^k)$  sei eine Folge von Abstiegsrichtungen erster Ordnung, und  $(t^k)$  sei eine effiziente Schrittweitenfolge. Dann gilt (2.6):

$$\lim_k \frac{\langle \nabla f(x^k), d^k \rangle}{\|d^k\|_2} = 0.$$

**Definition 2.2.7** (Gradientenbezogene Suchrichtungen). Die Folge von Suchrichtungen  $(d^k)$  heißt **gradientenbezogen**, falls folgendes gilt:

$$\exists c > 0 \ \forall k \in \mathbb{N} : \quad \frac{\langle \nabla f(x^k), d^k \rangle}{\|d^k\|_2} \leq -c \cdot \|\nabla f(x^k)\|_2$$

**Übung 2.2.8.** Die Suchrichtungen  $d^k = -\nabla f(x^k)$ ,  $k \in \mathbb{N}$  sind gradientenbezogen.

**Satz 2.2.9.** Die Menge  $f_{\leq}^{f(x^0)}$  sei beschränkt, und in Zeile 4 von Algorithmus 2.3 sei  $x^{k+1} = x^k + t^k d^k$  mit einer gradientenbezogenen Suchrichtungsfolge  $(d^k)$  und einer effizienten Schrittweitenfolge  $(t^k)$  gewählt. Für  $\epsilon = 0$  stoppt dann das Verfahren entweder nach endlich vielen Schritten mit einem kritischen Punkt, oder die Folge  $(x^k)$  besitzt einen Häufungspunkt, und für jeden solchen Punkt  $x^*$  gilt  $\nabla f(x^*) = 0$ .

**Korollar 2.2.10.** Die Menge  $f_{\leq}^{f(x^0)}$  sei beschränkt, und in Zeile 4 von Algorithmus 2.3 sei  $x^{k+1} = x^k + t^k d^k$  mit einer gradientenbezogenen Suchrichtungsfolge  $(d^k)$  und einer effizienten Schrittweitenfolge  $(t^k)$  gewählt. Dann terminiert das Verfahren für jedes  $\epsilon > 0$  nach endlich vielen Schritten.

## Schrittweitensteuerung

**Definition.** Eine Funktion  $F: D \rightarrow \mathbb{R}^m$  heißt **Lipschitz-stetig** auf  $D \subseteq \mathbb{R}^n$ , falls

$$\exists L > 0 \ \forall x, y \in D : \quad \|F(x) - F(y)\|_2 \leq L \cdot \|x - y\|_2$$

$C^1$ -Funktionen sind auf kompakten Mengen immer Lipschitz-stetig sind, damit ist  $\nabla f$  bei beschränkter Menge  $f_{\leq}^{f(x^0)}$  zum Beispiel für jede  $C^2$ -Funktion  $f$  Lipschitz-stetig auf  $f_{\leq}^{f(x^0)}$ .

**Bemerkung 2.2.12.** Bei beschränktem (und daher kompaktem)  $f_{\leq}^{f(x^0)}$  ist die Menge  $\text{conv}(f_{\leq}^{f(x^0)})$  ebenfalls kompakt, so dass die Forderung der Lipschitz-Stetigkeit von  $\nabla f$  auch auf  $\text{conv}(f_{\leq}^{f(x^0)})$  eine schwache Voraussetzung ist.

**Lemma 2.2.13.** Auf einer konvexen Menge  $D \subseteq \mathbb{R}^n$  sei  $f$  differenzierbar mit Lipschitz-stetigem Gradienten  $\nabla f$  und zugehöriger Lipschitz-Konstante  $L > 0$ . Dann gilt

$$\nabla \bar{x}, x \in D : \quad |f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle| \leq \frac{L}{2} \|x - \bar{x}\|_2^2$$

### Exakte Schrittweiten

Zu  $x \in f_{\leq}^{f(x^0)}$  sei eine Abstiegsrichtung erster Ordnung  $d$  für  $f$  in  $x$  gegeben. Wegen  $\varphi'_d(0) = \langle \nabla f(x), d \rangle < 0$  gilt  $\varphi_d(t) < \varphi_d(0)$  für kleine positive  $t$ . Für beschränktes  $f_{\leq}^{f(x^0)}$  besitzt  $\varphi_d$  nach dem Satz von Weierstraß sogar globale Minimalpunkte  $t_e > 0$ , die exakte Schrittweiten genannt werden. Per Definition der eindimensionalen Einschränkung  $\varphi_d$  erfüllen sie

$$f(x + t_e d) = \min_{t > 0} f(x + t d)$$

Eine exakte Schrittweite zu berechnen, um den größtmöglichen Abstieg von  $x$  aus entlang  $d$  zu erzielen, ist im Allgemeinen sehr aufwendig, so dass wir dieses Konzept meist nur für theoretische Zwecke benutzen werden und später stattdessen zu inexakten Schrittweiten übergehen werden. Es gilt

$$0 = \varphi_d(t_e) = \langle \nabla f(x + t_e d), d \rangle$$

**Übung 2.2.14.** Gegeben sei die quadratische Funktion  $q(x) = \frac{1}{2}x^T A x + b^T x$  mit  $A = A^T \succ 0$  und  $b \in \mathbb{R}^n$ , die nach Übung 1.2.24 koerziv und nach Übung 2.1.43 konvex ist. Für jedes  $x \in \mathbb{R}^n$  und jede Abstiegsrichtung erster Ordnung  $d$  für  $q$  in  $x$  die exakte Schrittweite eindeutig bestimmt zu

$$t_e = -\frac{\langle Ax + b, d \rangle}{d^T A d}$$

**Satz 2.2.15.** Die Menge  $f_{\leq}^{f(x^0)}$  sei beschränkt, die Funktion  $\nabla f$  sei Lipschitz-stetig auf  $\text{conv}(f_{\leq}^{f(x^0)})$ , und  $(d^k)$  sei eine Folge von Abstiegsrichtung erster Ordnung. Dann ist jede Folge von exakten Schrittweiten  $(t_e^k)$  effizient.

### Konstante Schrittweiten

Falls die Funktion  $f$  keine besondere Struktur aufweist, lohnt sich der Aufwand nicht, in jedem Iterationsschritt eine exakte Schrittweite  $t_e^k$  zu berechnen. Daher benutzt man dann lieber inexakte Schrittweiten, die ebenfalls effizient, aber erheblich leichter zu berechnen sind.

Eine zunächst naheliegend erscheinende Möglichkeit dafür besteht darin

$$t_c^k = -\frac{\langle \nabla f(x^k), d^k \rangle}{L \cdot \|d^k\|_2^2}$$

Auch diese Schrittweite ist effizient, genauso wie die exakte. Im speziellen Fall  $d^k = -\nabla f(x^k)$  gilt sogar

$$t_c^k = \frac{1}{L}$$

## Armijo-Schrittweiten

Eine in modernen Implementierungen von Optimierungsverfahren sehr beliebte inexacte Schrittweitensteuerung geht auf eine Idee von Armijo zurück: Zu  $x \in f_{\leq}^{f(x^0)}$  seien  $d$  eine Abstiegsrichtung erster Ordnung und  $\sigma \in (0, 1)$ . Dann existiert ein  $t > 0$ , so dass für alle  $t \in (0, \hat{t})$  die Werte  $\varphi_d(t)$  unter der „nach oben gedrehten Tangente“  $\varphi_d(0) + t\sigma\varphi_d'(0)$  liegen, so dass also gilt:

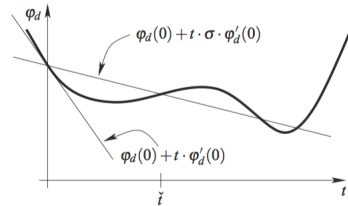
$$f(x + td) \leq f(x) + t\sigma \langle \nabla f(x), d \rangle$$

Offensichtlich erfüllt jedes solche  $t \in \mathbb{R}$  die Bedingung (2.3):

$$\exists c_1 > 0 \quad \forall k \in \mathbb{N} : \quad f(x^k + t^k d^k) - f(x^k) \leq c_1 \cdot t^k \langle \nabla f(x^k), d^k \rangle$$

mit  $c_1 = \sigma$ .

**Abb. 2.5** Armijo-Regel




---

### Algorithmus 2.4: Armijo-Regel

---

**Input :**  $C^1$ -Funktion  $f$  und  $x, d \in \mathbb{R}^n$  mit  $\langle \nabla f(x), d \rangle < 0$

**Output :** Armijo-Schrittweite  $t_a$

---

```

1 begin
2   Wähle  $\sigma, \rho \in (0, 1)$  sowie  $\gamma > 0$  (alle unabhängig von  $x$  und  $d$ ).
3   Wähle eine Startschrittweite  $t^0 \geq -\gamma \langle \nabla f(x), d \rangle / \|d\|_2^2$  und setze  $\ell = 0$ .
4   while  $f(x + t^\ell d) > f(x) + t^\ell \sigma \langle \nabla f(x), d \rangle$  do
5     Setze  $t^{\ell+1} = \rho t^\ell$ .
6     Ersetze  $\ell$  durch  $\ell + 1$ .
7   end
8   Setze  $t_a = t^\ell$ .
9 end
```

---

**Satz 2.2.16.** Die Menge  $f_{\leq}^{f(x^0)}$  sei beschränkt, die Funktion  $\nabla f$  sei Lipschitz-stetig auf  $\text{conv}(f_{\leq}^{f(x^0)})$ , und  $(d^k)$  sei eine Folge von Abstiegsrichtungen erster Ordnung. Dann ist die Folge der Armijo-Schrittweiten  $(t_a^k)$  aus Algorithmus 2.4 (mit unabhängig von  $k$  gewählten Parametern  $\sigma$ ,  $\rho$  und  $\gamma$ ) wohldefiniert und effizient.

**Übung 2.2.17.** Zeigen Sie für die Funktion  $f(x) = \frac{1}{2}x^2$ , den Startpunkt  $x^0 = -3$ , die Richtungen  $d^k = 2^{-k}$  sowie  $\sigma = \frac{1}{2}$ , dass der durch die Wahl  $t^0 := 1$  modifizierte Algorithmus 2.4 nicht zu einer effizienten Schrittweitenfolge führt.

Man sollte  $t^0$  also so initialisieren, wie in Algorithmus 2.4 angegeben, wobei sich die Wahl  $\gamma = 10^{-4}$  bewährt hat. Es ist außerdem nicht schwer zu sehen, dass sich die Armijo-Regel auch für nur einseitig richtungsdifferenzierbare Funktionen einsetzen lässt, indem man das Skalarprodukt  $\lambda \nabla f(\bar{x}), d \rangle$  durch  $f'(\bar{x}, d)$  ersetzt.

## 2.2.2 Gradientenverfahren

Aufgrund seiner geometrischen Grundidee ist dies das Verfahren des steilsten Abstiegs.

---

### Algorithmus 2.5: Gradientenverfahren

---

**Input :**  $C^1$ -Optimierungsproblem  $P$   
**Output :** Approximation  $\bar{x}$  eines kritischen Punkts von  $f$  (falls das Verfahren terminiert; Satz 2.2.18)

```

1 begin
2   Wähle einen Startpunkt  $x^0$ , eine Toleranz  $\varepsilon > 0$  und setze  $k = 0$ .
3   while  $\|\nabla f(x^k)\| > \varepsilon$  do
4     Setze  $d^k = -\nabla f(x^k)$ .
5     Bestimme eine Schrittweite  $t^k$ .
6     Setze  $x^{k+1} = x^k + t^k d^k$ .
7     Ersetze  $k$  durch  $k + 1$ .
8   end
9   Setze  $\bar{x} = x^k$ .
10 end
```

---

**Satz 2.2.18.** Die Menge  $f_{\leq}^{f(x^0)}$  sei beschränkt, die Funktion  $\nabla f$  sei Lipschitz-stetig auf  $\text{conv}(f_{\leq}^{f(x^0)})$ , und in Zeile 5 seien exakte Schrittweiten  $(t_e^k)$  oder Armijo-Schrittweiten  $(t_a^k)$  gewählt. Dann terminiert Algorithmus 2.5 für jedes  $\epsilon > 0$  nach endlich vielen Schritten. Falls eine Lipschitz-Konstante  $L > 0$  zur Lipschitz-Stetigkeit von  $\nabla f$  auf  $\text{conv}(f_{\leq}^{f(x^0)})$  bekannt ist, dann gilt dieses Ergebnis auch für die dann berechenbaren konstanten Schrittweiten  $t_c^k = L^{-1}, k \in \mathbb{N}$ .

**Definition.** Es ist  $\|A\|_2 := \max \{ \|Ad\|_2 \mid \|d\|_2 = 1 \}$ .

**Übung 2.2.19.** Gegeben sei die quadratische Funktion  $q(x) = \frac{1}{2}x^T Ax + b^T x$  mit  $A = A^T$  und  $b \in \mathbb{R}^n$ . Der Gradient  $\nabla q$  ist auf ganz  $\mathbb{R}^n$  Lipschitz-stetig mit  $L = \|A\|_2$ .

**Beispiel 2.2.20.** Nach Übung 2.2.19 erzeugt das Gradientenverfahren eine sogar gegen den globalen Minimalpunkt von  $q$  konvergente Folge von Iterierten  $(x^k)$ , wenn entweder exakte, konstante oder Armijo-Schrittweiten gewählt werden.

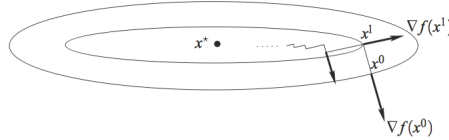
Nach Übung 2.2.14 ist bei jeder Abstiegsrichtung erster Ordnung für  $q$  in  $x$  die (eindeutige)

exakte Schrittweite beim Gradientenverfahren

$$t_e = \frac{\|\nabla q(x)\|_2^2}{Dq(x)A\nabla q(x)}$$

Falls die Höhenlinien von  $f$  die Form lang gezogener Ellipsen mit einem Minimalpunkt  $x^*$  in deren gemeinsamem Zentrum besitzen, dann zeigt  $-\nabla f(x^k)$  typischerweise nicht in die Richtung von  $x^*$ . Die Iterierten springen dadurch entlang einer Zickzacklinie, weshalb man in Anlehnung an die englischsprachige Literatur auch vom Zigzagging-Effekt spricht.

**Abb. 2.6** Zigzagging-Effekt



**Definition 2.2.21** (Konvergenzgeschwindigkeiten). Es sei  $(x^k)$  eine konvergente Folge mit Grenzpunkt  $x^*$ . Sie heißt

a) **linear konvergent**, falls  $\exists 0 < c < 1, k_0 \in \mathbb{N} \forall k \geq k_0$ :

$$\|x^{k+1} - x^*\| \leq c \cdot \|x^k - x^*\|,$$

b) **superlinear konvergent**, falls  $\exists c^k \searrow 0, k_0 \in \mathbb{N} \forall k \geq k_0$ :

$$\|x^{k+1} - x^*\| \leq c^k \cdot \|x^k - x^*\|,$$

c) **quadratisch konvergent**, falls  $\exists c > 0, k_0 \in \mathbb{N} \forall k \geq k_0$ :

$$\|x^{k+1} - x^*\| \leq c \cdot \|x^k - x^*\|^2.$$

Der folgende Satz zeigt, dass das Gradientenverfahren schon für sehr angenehme Funktionen nur linear konvergente Funktionswerte der Iterierten besitzt, und zwar mit einer Konstante  $c$ , die sehr nahe bei eins liegen kann. Konkret betrachten wir die konvex-quadratische Funktion  $q(x) = \frac{1}{2}x^T A x + b^T x$  mit  $A = A^T \succ 0$  sowie  $b \in \mathbb{R}^n$  und bezeichnen den größten und den kleinsten Eigenwert der Matrix  $A$  mit  $\lambda_{\max}$  bzw.  $\lambda_{\min}$  - nach Beispiel 2.2.20 konvergieren dabei die Iterierten des Gradientenverfahrens mit exakten Schrittweiten gegen den globalen Minimalpunkt  $x = -A^{-1}b$  von  $q$ .

**Lemma 2.2.22** (Kantorowitsch-Ungleichung). Es sei  $A = A^T \succ 0$  mit maximalem und minimalem Eigenwert  $\lambda_{\max}$  bzw.  $\lambda_{\min}$ . Dann gilt für jedes  $v \in \mathbb{R}^n \setminus \{0\}$

$$\frac{v^T A^{-1} v \cdot v^T A v}{\|v\|_2^4} \leq \frac{(\lambda_{\max} + \lambda_{\min})^2}{4\lambda_{\max}\lambda_{\min}}$$

**Satz 2.2.23.** Auf die konvex-quadratische Funktion  $q(x) = \frac{1}{2}x^T Ax + b^T x$  mit  $A = A^T \succ 0$  und  $b \in \mathbb{R}^n$  werde das Gradientenverfahren mit exakten Schrittweiten und  $\epsilon = 0$  angewendet. Dann gilt für alle  $k \in \mathbb{N}$

$$|q(x^{k+1}) - q(x^*)| \leq \left( \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 |q(x^k) - q(x^*)|.$$

Nach Satz 2.2.23 minimiert das Gradientenverfahren (mit exakten Schrittweiten) eine konvex-quadratische Funktion  $q(x) = \frac{1}{2}x^T Ax + b^T x$  in einem einzigen Schritt, wenn der kleinste und größte Eigenwert  $\lambda_{\min}$  bzw.  $\lambda_{\max}$  von  $A$  übereinstimmen. Dann stimmen natürlich auch alle Eigenwerte von  $A$  miteinander überein, so dass  $q$  sphärenförmige Niveaumengen besitzt.

## Variable-Metrik-Verfahren

Im Allgemeinen existiert stets ein rechtwinkliges Koordinatensystem, das zur Lage der ellipsodialen Niveaumengen von  $q$  „passend ausgerichtet“ ist, sodass in einem neuen Koordinatensystem die Niveaulinien sphärenförmig sind.

Für eine nicht notwendigerweise konvex-quadratische Funktion  $f \in C^1(\mathbb{R}^n, \mathbb{R})$  beschränkt man sich darauf, approximativ eine Konstruktion wie bei konvex-quadratischen Funktionen zu benutzen:

**Definition 2.2.27** (Gradient bezüglich einer positiv definiten Matrix). Für  $f \in C^1(\mathbb{R}^n, \mathbb{R})$  und eine  $(n, n)$ -Matrix  $A = A^T \succ 0$  heißt

$$\nabla_A f(x) := A^{-1} \nabla f(x)$$

**Gradient von  $f$  bezüglich  $A$  an  $x$ .**

Die verschiedenen Variable-Metrik-Verfahren unterscheiden sich durch die Wahl der Matrix  $A$ , mit deren Hilfe die Suchrichtung  $-\nabla_A f(x)$  gebildet wird. Für jedes  $A = A^T \succ 0$  ist diese Suchrichtung an einem nichtkritischen Punkt  $x$  jedenfalls eine Abstiegsrichtung erster Ordnung, denn da mit  $A$  auch  $A^{-1}$  positiv definit ist, gilt

$$\langle \nabla f(\bar{x}), -\nabla_A f(\bar{x}) \rangle = -\nabla f(\bar{x})^T A^{-1} \nabla f(\bar{x}) < 0$$

**Übung 2.2.28.** Für jedes  $A = A^T \succ 0$  ist die Funktion  $\langle x, y \rangle_A := x^T A y$  ein Skalarprodukt auf  $\mathbb{R}^n$ .

**Übung 2.2.29.** Für jedes  $A = A^T \succ 0$  und für das von  $A$  induzierte Skalarprodukt  $\langle \cdot, \cdot \rangle_A$  ist die folgende Funktion eine Norm auf  $\mathbb{R}^n$ :

$$\|x\|_A := \sqrt{\langle x, x \rangle_A}$$



**Übung 2.2.30.** Es gilt unter für die konvex-quadratische Funktion  $q(x) = \frac{1}{2}x^T Ax + b^T x$  mit  $A = A^T \succ 0$ ,  $b \in \mathbb{R}^n$  und mit exakten Schrittweiten:

$$\frac{1}{2}\|x - x^*\|_A^2 = q(x) - q(x^*)$$

Da die quadrierten Abstände der Iterierten zum Grenzpunkt linear konvergieren, erhält man aus Übung 2.2.30 eine sogar noch langsamere als lineare (nämlich eine sog. sublineare) Konvergenzgeschwindigkeit der Iterierten selbst.

**Übung 2.2.31.** Gegeben sei die quadratische Funktion  $q(x) = \frac{1}{2}x^T Ax + b^T x$  mit  $A = A^T$  und  $b \in \mathbb{R}^n$ . Für die exakte Schrittweite des Gradientenverfahrens gilt die Formel

$$t_e = \frac{\|\nabla q(x)\|_2^2}{\|\nabla q(x)\|_A^2}$$

**Übung 2.2.32.** Für das durch  $A = A^T \succ 0$  induzierte Skalarprodukt  $\langle \cdot, \cdot \rangle_A$  und die induzierte Norm  $\|\cdot\|_A$  gilt die Cauchy-Schwarz-Ungleichung:

$$\forall x, y \in \mathbb{R}^n : \quad |\langle x, y \rangle_A| \leq \|x\|_A \cdot \|y\|_A$$

und die Abschätzung ist scharf.

**Lemma 2.2.33.** Es sei  $\nabla f(x) \neq 0$ . Dann löst der Vektor

$$d = -\frac{\nabla_A f(x)}{\|\nabla_A f(x)\|_A}$$

das Problem  $\min \langle \nabla f(x), d \rangle$  s.t.  $\|d\|_A = 1$ , und zwar mit optimalem Wert  $-\|\nabla_A f(x)\|_A$ .

---

**Algorithmus 2.6:** Variable-Metrik-Verfahren

---

**Input :**  $C^1$ -Optimierungsproblem  $P$

**Output :** Approximation  $\bar{x}$  eines kritischen Punkts von  $f$  (falls das Verfahren terminiert; [Satz 2.2.37](#))

---

```

1 begin
2   Wähle einen Startpunkt  $x^0$ , eine Matrix  $A^0 = (A^0)^T \succ 0$ , eine Toleranz  $\varepsilon > 0$  und
   setze  $k = 0$ .
3   while  $\|\nabla f(x^k)\|_2 > \varepsilon$  do
4     Setze  $d^k = -\nabla_{A^k} f(x^k)$ .
5     Bestimme eine Schrittweite  $t^k$ .
6     Setze  $x^{k+1} = x^k + t^k d^k$ .
7     Wähle  $A^{k+1} = (A^{k+1})^T \succ 0$ .
8     Ersetze  $k$  durch  $k + 1$ .
9   end
10  Setze  $\bar{x} = x^k$ .
11 end
```

---

In Zeile 2 von Algorithmus 2.6 wählt man häufig  $A^0 = E$ , also als erste Suchrichtung die Gradientenrichtung  $d^0 = -\nabla f(x^0)$ . In Zeile 3 wäre ein konsistenteres Abbruchkriterium eigentlich

$$\|\nabla_{A^k} f(x^k)\|_{A^k} \leq \epsilon,$$

aber wegen

$$\|\nabla_{A^k} f(x^k)\|_{A^k} = \sqrt{Df(x^k)(A^k)^{-1}\nabla f(x^k)} = \|\nabla f(x^k)\|_{(A^k)^{-1}}$$

und der Äquivalenz von  $\|\cdot\|_{(A^k)^{-1}}$  und  $\|\cdot\|_2$  (d. h., es gibt Konstanten  $c_1, c_2 > 0$ , so dass alle  $x \in \mathbb{R}^n$  die Abschätzungen  $c_1\|x\|_{(A^k)^{-1}} \leq \|x\|_2 \leq c_2\|x\|_{(A^k)^{-1}}$  erfüllen) kann man ebensogut das angegebene und weniger aufwendigere Kriterium testen. In Zeile 4 berechnet man die Suchrichtung  $d^k$  numerisch nicht durch die eine Matrixinversion enthaltende Definition  $-(A^k)^{-1}\nabla f(x^k)$ , sondern weniger aufwendig als Lösung des linearen Gleichungssystems  $A^k d = -\nabla f(x^k)$ .

Möchte man die Konvergenz von Variable-Metrik-Verfahren im Sinne von Satz 2.2.9 garantieren, benötigt man neben der Effizienz der Schrittweiten auch die Gradientenbezogenheit der Suchrichtungen. Diese muss man noch fordern.

**Definition 2.2.34** (Gleichmäßig positiv definite und beschränkte Matrizen). Eine Folge  $(A^k)$  symmetrischer  $(n, n)$ -Matrizen heißt **gleichmäßig positiv definit** und **beschränkt**, falls folgendes gilt:

$$\exists 0 < c_1 \leq c_2 \forall d \in B_=(0, 1), k \in \mathbb{N} : \quad c_1 \leq d^T A^k d \leq c_2$$

**Übung 2.2.35.** Die Folge  $(A^k)$  sei gleichmäßig positiv definit und beschränkt mit Konstanten  $c_1$  und  $c_2$ . Dann ist die Folge  $((A^k)^{-1})$  auch gleichmäßig positiv definit und beschränkt mit Konstanten  $\frac{1}{c_2}$  und  $\frac{1}{c_1}$ . Außerdem ist die Folge  $(\lambda_{\max}((A^k)^{-1}))$  der größten Eigenwerte von  $((A^k)^{-1})$  durch  $\frac{1}{c_1}$  nach oben beschränkt.

**Satz 2.2.36.** Die Folge  $(A^k)$  sei gleichmäßig positivdefinit und beschränkt. Dann ist die Folge  $(d^k)$  mit  $d^k = -(A^k)^{-1}\nabla f(x^k)$ ,  $k \in \mathbb{N}$ , gradientenbezogen.

**Satz 2.2.37.** Die Menge  $f_{\leq}^{f(x^0)}$  sei beschränkt, die Funktion  $\nabla f$  sei Lipschitz-stetig auf  $\text{conv}(f_{\leq}^{f(x^0)})$ , die Folge  $(A^k)$  sei gleichmäßig positiv definit und beschränkt, und in Zeile 5 seien exakte Schrittweite  $(t_e^k)$  oder Armijo-Schrittweiten  $(t_a^k)$  gewählt. Dann terminiert Algorithmus 2.6 für jedes  $\epsilon > 0$  nach endlich vielen Schritten.

**Bemerkung 2.2.38** (Spektralnrm und Eigenwerte). Durch die in Abschnitt 2.2.3 eingeführte Spektralnrm  $\|A\|_2 := \max \{\|Ad\|_2 \mid \|d\|_2 = 1\}$  können die Eigenwerte einer Matrix  $A$  durch die semidefinite Matrix  $A^T A$  berechnet werden. Es ist

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sqrt{(\lambda_{\max}(A))^2} = |\lambda_{\max}(A)|$$

Damit ist nach Bemerkung 2.2.24 und Übung 2.2.5 die Länge der längsten Halbachse des Niveau-Ellipsoids

$$\frac{1}{\sqrt{\lambda_{\min}(A^T A)^{-1}}} = \sqrt{\lambda_{\max}(A^T A)}$$

## Newton-Verfahren mit und ohne Dämpfung

Wählt man in Algorithmus 2.6 für  $f \in C^2(\mathbb{R}^n, \mathbb{R})$  in jedem Schritt  $A^k = D^2 f(x^k)$ , so erhält man das Newton-Verfahren, sofern die Matrizen  $D^2 f(x^k)$  positiv definit sind.

Ist  $x^*$  nichtdegenerierter lokaler Minimalpunkt von  $f$ , dann gilt aus den bereits am Ende von Abschnitt 2.1.5. aufgeführten Stetigkeitsgründen  $D^2 f(x) \succ 0$  für alle  $x$  aus einer Umgebung von  $x^*$ . Für  $x^0$  aus dieser Umgebung kann man also  $A^k = D^2 f(x^k)$  setzen und erhält ein wohldefiniertes Abstiegsverfahren. Ferner sind die Suchrichtungen  $d^k = -(D^2 f(x^k))^{-1} \nabla f(x^k)$  gradientenbezogen, falls  $f$  und  $x$  gleichmäßig konvex ist, d. h. falls für eine Umgebung  $U$  von  $x$  gilt:

$$\exists c > 0 \forall x \in U, d \in B_=(0, 1) : \quad c \leq d^T D^2 f(x) d$$

Die für diese Folgerung nach Satz 2.2.36 noch erforderliche Beschränktheit der Folge  $(D^2 f(x^k))$  resultiert dabei aus der Stetigkeit von  $D^2 f$ . Die Nichtdegeneriertheit des lokalen Minimalpunkts  $x^*$  gilt bei gleichmäßig konvexem  $f$  automatisch.

Die Dämpfung des Newton-Verfahrens hat den Vorteil, dass der Konvergenzradius (also der mögliche Abstand von  $x_0$  zu  $x^*$ ) etwas größer wird. Andererseits ist zunächst nicht klar, ob die Dämpfung nicht auch die lokale Konvergenz verlangsamt. Das ungedämpfte Newton-Verfahren konvergiert unter schwachen Voraussetzungen jedenfalls quadratisch.

**Satz 2.2.39** (Quadratische Konvergenz des Newton-Verfahrens). Die durch

$$x^{k+1} = x^k - \left(D^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

definierte Folge  $(x^k)$  konvergiere gegen einen nichtdegenerierten lokalen Minimalpunkt  $x^*$ , und  $D^2 f$  sei Lipschitz-stetig auf einer konvexen Umgebung von  $x^*$ . Dann konvergiert die Folge  $(x^k)$  quadratisch gegen  $x^*$ .

**Bemerkung 2.2.40.** Die Voraussetzungen von Satz 2.2.39 lassen sich noch erheblich abschwächen. Erstens gilt die Aussage für jeden nichtdegenerierten kritischen Punkt  $x^*$  also nicht nur für lokale Minimalpunkte. Zweitens zeigt Satz 2.2.48, dass die Konvergenz der Folge  $(x^k)$  bereits impliziert, dass der Grenzpunkt  $x^*$  ein nichtdegenerierter kritischer Punkt ist.

Die Konvergenzgeschwindigkeit überträgt sich aus Satz 2.2.39 natürlich auf das gedämpfte Newton-Verfahren, falls man mit einem  $k_0 \in \mathbb{N}$  für alle  $k \geq k_0$  nur  $t_k = 1$  wählt. Die folgende Übung gibt eine natürliche Bedingung dafür an.

**Übung 2.2.41.** Für  $f \in C^2(\mathbb{R}^n, \mathbb{R})$  liege  $x$  in einer genügend kleinen Umgebung eines nichtdegenerierten lokalen Minimalpunkts, und die Suchrichtung  $d$  werde mit dem gedämpften Newton-Verfahren per Armijo-Regel mit  $t^0 = 1$  und  $\sigma < \frac{1}{2}$  bestimmt. Dann gilt  $\langle \nabla f(x), d \rangle < 0$  und dass die Armijo-Regel die Schrittweite  $t_a = 1$  wählt.

**Übung 2.2.42.** Das ungedämpfte Newton-Verfahren liefert für die Funktion  $q(x) = \frac{1}{2}x^T Ax + b^T x$  mit  $A = A^T \succ 0$  und  $b \in \mathbb{R}^n$  von jedem Startpunkt  $x^0 \in \mathbb{R}^n$  nach einem Schritt den globalen Minimalpunkt von  $q$ .

**Übung 2.2.43.** Für  $f \in C^2(\mathbb{R}^n, \mathbb{R})$  sei eine Iterierte  $x^k$  mit  $D^2 f(x^k) \succ 0$  gegeben. Die vom Newton-Verfahren erzeugte Suchrichtung  $d^k$  ist der eindeutige lokale Minimalpunkt der konvex-quadratischen Funktion

$$q(d) = f(x^k) + \langle \nabla f(x^k), d \rangle + \frac{1}{2} d^T D^2 f(x^k) d$$

**Bemerkung.** Für spezielle Funktionen, wie bei Kleinste-Quadrate-Problemen, also  $f(x) = \frac{1}{2} \|r(x)\|_2^2$  mit einer glatten Funktion  $r: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . Es ist

$$D^2 f(x) = \nabla r(x) D r(x) + \sum_{j=1}^m r_j(x) D^2 r_j(x).$$

Als Modifikation, dem Gauß-Newton-Verfahren, wäre anstatt  $A^k = D^2 f(x^k)$  in Zeile 4  $A^k = \nabla r(x^k) D r(x^k)$  zu wählen. Dass die restlichen Summanden in der Darstellung von  $A^k$  eine untergeordnete Rolle spielen, kann zum einen daran liegen, dass für  $m \geq n$  üblicherweise ein Punkt  $x^*$  mit  $r(x^*) = 0$  existiert, so dass für  $m \geq n$  approximiert wird, so dass die Werte  $r_j(x^k)$  fast verschwinden, oder zum anderen daran, dass die Krümmung der Funktionen  $r_j$  an  $x^*$  vernachlässigbar sind, so dass sich die Matrizen  $D^2 r_j(x^k)$  in der Nähe der Nullmatrix aufhalten.

Obwohl zum Aufstellen von  $A^k$  im Gauß-Newton-Verfahren also nur Ableitungsinformationen erster Ordnung (die Matrix  $D r(x^k)$ ) erforderlich sind, lässt sich unter bestimmten Zusatzvoraussetzungen sogar quadratische Konvergenz zeigen. Zusätzlich sind die Suchrichtungen  $d^k$  im Gauß-Newton-Verfahren im Gegensatz zum allgemeinen Newton-Verfahren garantiert Abstiegsrichtungen (erster Ordnung), so dass eine Schrittweitensteuerung etwa per Armijo-Regel möglich ist.

Sollte die Jacobi-Matrix  $D r(x^k)$  nicht den vollen Rang besitzen oder zumindest schlecht konditioniert sein, so lässt das Gauß-Newton-Verfahren sich durch die Wahl  $A^k = \nabla r(x^k) D r(x^k) + \sigma^k E$  mit gewissen  $\sigma^k > 0$  und der Einheitsmatrix  $E$  passender Dimension stabilisieren, was auf das Levenberg-Marquardt-Verfahren führt.

## Superlineare Konvergenz

Falls im Newton-Verfahren  $x^0$  zu weit von einem nichtdegenerierten Minimalpunkt entfernt liegt, ist  $D^2 f(x^k)$  nicht notwendigerweise positiv definit und die Newton-Richtung entweder

nicht definiert oder nicht notwendigerweise eine Abstiegsrichtung. Man versucht daher, das Newton-Verfahren zu globalisieren, d. h. Konvergenz im Sinne von Satz 2.2.9 gegen einen lokalen Minimalpunkt von jedem Startpunkt  $x^0 \in \mathbb{R}^n$  aus zu erzwingen. Ein erster Ansatz dazu besteht darin, in Zeile 2 von Algorithmus 2.6  $A^0 = E$  zu wählen sowie in Zeile 7 (ähnlich wie im Levenberg-Marquardt-Verfahren)

$$A^{k+1} = D^2 f(x^{k+1}) + \sigma^{k+1} \cdot E$$

mit einem so großen Skalar  $\sigma^{k+1}$ , dass  $A^{k+1}$  positiv definit ist, und bei hinreichend großen  $k$  wieder  $\sigma^k = 0$  (d. h. das Verfahren startet als Gradientenverfahren und geht nach endlich vielen Schritten in das gedämpfte Newton-Verfahren über). Ein Nachteil des Verfahrens besteht darin, dass die Bestimmung von  $\sigma_k$  sehr aufwendig sein kann: Man halbiert oder verdoppelt  $\sigma^k$  so lange, bis ein Test auf positive Definitheit von erfolgreich ist.

Im Folgenden werden wir Verfahren kennenlernen, die nicht nach endlich vielen Schritten, sondern nur asymptotisch in das gedämpfte Newton-Verfahren übergehen. Für diese lässt sich immerhin noch superlineare Konvergenz zeigen. Der entsprechende Konvergenzsatz erfordert einige Vorbereitungen.

Zunächst besitzt die Folge der Iterierten  $(x^k)$  nach Satz 2.2.9 einen Häufungspunkt, und jeder solche Häufungspunkt ist kritisch, sofern die Menge  $f_{\leq}^{f(x^0)}$  beschränkt ist und gradientenbezogene Suchrichtungen sowie effiziente Schrittweiten benutzt werden. Die Gradientenbezogenheit der Suchrichtungen wird durch Satz 2.2.36 für gleichmäßig positiv definite und beschränkte  $(A^k)$  garantiert. Sei dazu

$$H^k := t^k \left( A^k \right)^{-1},$$

sodass  $x^{k+1} = x^k - H^k \nabla f(x^k)$ . Daraus folgt die Definition der superlinearen Konvergenz und diese ist äquivalent zu  $\limsup_k \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0$ .

**Lemma 2.2.46.** Die Folge  $(x^k)$  sei nach obiger Vorschrift gebildet und gegen  $x^*$  konvergent. Ferner seien die Folgen  $(\|H^k\|_2)$  und  $(\|(H^k)^{-1}\|_2)$  beschränkt. Dann gilt

- a)  $\nabla f(x^*) = 0$
- b)  $\limsup_k \|x^{k+1} - x^*\|_2 / \|x^k - x^*\|_2 \leq \limsup_k \|E - H^k D^2 f(x^*)\|_2$

**Lemma 2.2.47.** Für zwei  $(n, n)$ -Matrizen  $A$  und  $B$  sei  $L := \|E - AB\|_2 < 1$ . Dann gilt

- a)  $A$  und  $B$  sind nichtsingulär
- b)  $\|A\|_2 \leq (1 + L) \cdot \|B^{-1}\|_2$
- c)  $\|A^{-1}\| \leq \frac{\|B\|_2}{(1-L)}$

**Satz 2.2.48.** Die Folge  $(x^k)$  sei nach obiger Vorschrift gebildet und gegen  $x^*$  konvergent. Ferner sei  $L := \limsup_k \|E - H^k D^2 f(x^*)\|_2 < 1$ . Dann gelten die folgenden Aussagen:

- a)  $D^2 f(x^*)$  ist nichtsingulär.
- b)  $\nabla f(x^*) = 0$
- c)  $(x^k)$  konvergiert mindestens linear gegen  $x^*$
- d) Es gilt  $L = 0$  genau im Fall von  $\lim_k H^k = (D^2 f(x^*))^{-1}$ , und in diesem Fall konvergiert  $(x^k)$  superlinear gegen  $x^*$

Nach Satz 2.2.48 sollte Algorithmus 2.6 also asymptotisch in das ungedämpfte Newton-Verfahren übergehen, um superlineare Konvergenz zu garantieren. Wegen

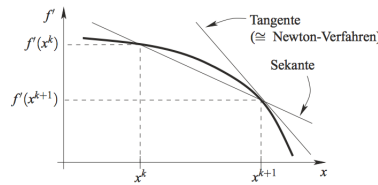
$$H^k = t^k \cdot (A^k)^{-1}$$

sind natürliche Bedingungen dafür  $\lim_k t^k = 1$  und  $\lim_k A^k = D^2 f(x^*)$ . Das zu Beginn dieses Kapitels vorgeschlagene Verfahren erreicht dies mit hohem Aufwand bereits nach endlich vielen Schritten, ist in diesem Sinne also nicht effizient.

## Quasi-Newton-Verfahren

Ein möglicher Ansatz Matrizen  $A^k$  zu finden mit  $\lim_k A^k = D^2 f(x^*)$  besteht darin, zunächst das Sekantenverfahren zur Nullstellensuche einer Funktion von  $\mathbb{R}$  nach  $\mathbb{R}$  zu betrachten.

**Abb. 2.8** Grundidee des Sekantenverfahrens



Hieraus erhält man für  $n \geq 1$  die Gleichung

$$\nabla f(x^{k+1}) - \nabla f(x^k) = A^{k+1} \cdot (x^{k+1} - x^k)$$

die als Sekantengleichung oder Quasi-Newton-Bedingung an die  $(n, n)$ -Matrix  $A^{k+1}$  bezeichnet wird. Man zählt leicht nach, dass (2.13)  $n$  Gleichungen für die  $n^2$  Einträge von  $A^{k+1}$  liefert. Selbst wenn man  $A^{k+1}$  als symmetrisch voraussetzt, sind noch immer  $n(n+1)/2$  Einträge zu bestimmen. Aus diesem Grunde existieren viele Möglichkeiten, verschiedene Quasi-Newton-Verfahren anzugeben.

Die Grundidee der folgenden Verfahren besteht darin, die Matrix  $A^{k+1}$  nicht in jedem Iterationsschritt komplett neu zu berechnen, sondern sie als möglichst einfaches Update

der Matrix  $A^k$  aus dem vorherigen Schritt aufzufassen. Als erfolgreicher Ansatz hat sich dabei erwiesen, mit  $A^0 \succ 0$  zu starten und in Zeile 7 von Algorithmus 2.6 die Matrix  $A^{k+1}$  aus  $A^k$  durch Addition einer symmetrischen Matrix vom Rang eins oder zwei zu gewinnen:

$$A^{k+1} = A^k + \alpha_k(u^k)(u^k)^T + \beta_k(v^k)(v^k)^T$$

mit Skalaren  $\alpha_k, \beta_k \in \mathbb{R}$  und Vektoren  $u^k, v^k \in \mathbb{R}^n$ , die so gewählt sind, dass  $A^{k+1}$  die Sekantengleichung erfüllt. Mit  $s^k := x^{k+1} - x^k$  und  $y^k := \nabla f(x^{k+1}) - \nabla f(x^k)$  heißt

$$A^{k+1} = A + \frac{yy^T}{y^T s} - \frac{Ass^T A}{s^T A s}$$

BFGS-Update. Da man allerdings in Zeile 4 von Algorithmus 2.6 die Suchrichtung

$$d^k = -\left(A^k\right)^{-1} \nabla f(x^k)$$

wählt, wäre es günstig, die Matrix  $(A^k)^{-1}$  explizit angeben zu können.

**Übung 2.2.49** (Sherman-Morrison-Woodbury-Formel).

- Für eine nichtsinguläre  $(n, n)$ -Matrix  $A$  und Vektoren  $b, c \in \mathbb{R}^n$  gelte, dass  $A + bc^T$  genau dann nichtsingulär ist, wenn  $1 + c^T A^{-1} b$  nicht verschwindet.
- Es gilt die Sherman-Morrison-Woodbury-Formel für eine  $(n, n)$ -Matrix  $A$  und Vektoren  $b, c \in \mathbb{R}^n$ , wobei  $A$  und  $A + bc^T$  nichtsingulär seien:

$$(A + bc^T)^{-1} = A^{-1} - \frac{A^{-1}bc^T A^{-1}}{1 + c^T A^{-1} b}$$

Übung 2.2.49 liefert eine Update-Formel für die inversen Matrizen  $B := A^{-1}$  und  $B^+ := (A^+)^{-1}$ , nämlich

$$B_{BFGS}^+ = B + \frac{ss^T}{s^T y} - \frac{Byy^T B}{y^T B y} + rr^T$$

mit  $r := \sqrt{y^T B y} \cdot \left( \frac{s}{s^T y} - \frac{By}{y^T B y} \right)$ . In Zeile 2 wählt man daher eine Matrix  $B^0 \succ 0$  anstelle von  $A^0$ . In Zeile 4 setzt man

$$d^k = -B^k \cdot \nabla f(x^k)$$

und in Zeile 7 wählt man  $B^{k+1} = B_{BFGS}^{k+1}$ . Analog kann man auch das DFP-Update definieren, und zwar mit  $s$  und  $y$  vertauscht

$$D_{DFP}^+ = B + \frac{ss^T}{s^T y} - \frac{Byy^T B}{y^T B y}$$

Dies unterscheidet sich vom BFGS-Update lediglich durch den Term  $rr^T$ . Die Einführung eines zusätzlichen Parameters  $\theta \in \mathbb{R}$  liefert die Updates der Broyden-Familie

$$B_\theta^+ = B_{DFP}^+ + \theta \cdot rr^T$$

Offenbar gilt  $B_0^+ = B_{DFP}^+$  und  $B_1^+ = B_{BFGS}^+$ . Außerdem gilt für die Wahl  $\theta = \frac{s^T y}{s^T y - y^T B y}$ :

$$B_{SR1}^+ := B_\theta^+ = B + \frac{(s - By)(s - By)^T}{(s - By)^T y}$$

so dass die Update-Matrix nur den Rang 1 besitzt. Man spricht dann vom SR1-Update (SR1 = symmetric rank 1).

**Übung 2.2.50.** Sei  $B^k \succ 0$  in einer Iteration eines Quasi-Newton-Verfahrens. Unter der Wahl von exakten Schrittweiten  $t_e^k$  gilt die Ungleichung

$$(y^k)^T s^k > 0$$

**Lemma 2.2.51.** Es sei  $\theta \geq 0$  beliebig. Dann gilt unter den Bedingungen  $B \succ 0$  und  $s^T y > 0$  auch  $B_\theta^+ \succ 0$

Zur Division durch die Zahlen  $s^T y$  und  $y^T B y$  in den Update-Formeln lässt sich also anmerken, dass für mit  $B \succ 0$  auch alle iterierten Matrizen  $B^k$  positiv definit sind, sofern  $(s^k)^T y^k > 0$  positiv ist. Insbesondere gilt dann  $y_k \neq 0$  und  $(y^k)^T B^k y^k > 0$ . Angemerkt sei, dass die Voraussetzung  $\theta$  aus Lemma 2.2.51 zwar für den SR1-Update nicht garantiert ist, er in der Praxis aber dennoch häufig gute Ergebnisse liefert.

Wählt man exakte Schrittweiten  $t_e^k$ , so hängt nur der Koeffizient des Vektors  $r^k$  von  $\theta$  ab, und somit ist die Suchrichtung für jedes  $\theta \in \mathbb{R}$  identisch. Weil man aber entlang dieser Richtung exakt eindimensional minimiert, liefern alle Verfahren der Broyden-Familie identische Lösungsfolgen  $(x^k)$ .

Dieses überraschende Ergebnis wird dadurch relativiert, dass man in der Praxis meist nicht exakt, sondern inexakt eindimensional minimiert, etwa per Armijo-Schrittweitensteuerung mit Backtracking Line Search. Während zum Beispiel das DFP-Update dazu tendiert, schlecht konditionierte Matrizen  $B^k$  zu erzeugen, verhält sich das BFGS-Update für Probleme mittlerer Größe numerisch oft sehr robust.

Leider lässt sich nicht zeigen, dass die Matrizen  $B^k$  stets gegen  $(D^2 f(x^*))^{-1}$  streben, wie es zur Anwendung von Satz 2.2.48 zur superlinearen Konvergenz wünschenswert wäre. Mit einer recht technischen Verallgemeinerung von Satz 2.2.48 lässt sich für  $\lim_k t^k = 1$  trotzdem die superlineare Konvergenz der BFGS- und DFP-Verfahren nachweisen, falls  $(B^k)^{-1}$  und  $D^2 f(x^*)$  wenigstens entlang der Suchrichtungen  $d^k$  asymptotisch gleich sind.

## Konjugierte Richtungen

Für viele Anwendungsprobleme ist die Anzahl der Variablen so hoch, dass sich zwar Vektoren der Länge  $n$  wie  $x^k$  und  $d^k$  noch gut abspeichern lassen, die Speicherung der  $n(n+1)/2$  Einträge von Matrizen wie  $B^k$  aber zu einem Platzproblem führt.



**Definition 2.2.52** (Konjugiertheit bezüglich einer positiv definiten Matrix). Es sei  $A$  eine  $(n, n)$ -Matrix mit  $A = A^T \succ 0$ . Zwei Vektoren  $v, w \in \mathbb{R}^n$  heißen **konjugiert bezüglich**  $A$ , falls  $\langle v, w \rangle = 0$  gilt.

Im Folgenden betrachten wir das allgemeine Abstiegsverfahren  $x^{k+1} = x^k + t_e^k d^k$  mit exakten Schrittweiten  $t_e^k$  und Abstiegsrichtungen erster Ordnung  $d^k$  für die konvex-quadratische Funktion

$$q(x) = \frac{1}{2} x^T A x + b^T x$$

mit  $A = A^T \succ 0$  und  $b \in \mathbb{R}^n$ .

**Übung 2.2.53.** Für  $k \in \mathbb{N}$  seien  $d^0, \dots, d^k$  paarweise konjugiert bezüglich  $A$  und sämtlich ungleich null. Es gilt

- a) Die Vektoren  $d^0, \dots, d^k$  sind linear unabhängig. insbesondere gilt  $k < n$
- b) Für  $k = n - 1$  gilt

$$A^{-1} = \sum_{l=0}^{n-1} \frac{(d^l)(d^l)^T}{(d^l)^T A (d^l)}$$

**Lemma 2.2.54.** Für  $k \in \mathbb{N}$  seien  $d^0, \dots, d^k$  paarweise konjugiert bezüglich  $A$ . Dann gilt

$$\forall 0 \leq l \leq k : \quad \langle \nabla q(x^{k+1}), d^l \rangle = 0.$$

**Satz 2.2.55.** Die Vektoren  $d^0, \dots, d^{n-1}$  seien paarweise konjugiert bezüglich  $A$  und sämtlich ungleich null. Dann ist  $x^n$  der globale Minimalpunkt von  $q$ .

Satz 2.2.55 besagt, dass ein Abstiegsverfahren für die konvex-quadratische Funktion  $q$  bei exakter Schrittweitensteuerung und paarweise konjugierten Suchrichtungen nach höchstens  $n$  Schritten den globalen Minimalpunkt von  $q$  findet. Da für ein Abstiegsverfahren wegen  $f(x^{k+1}) < f(x^k)$  stets  $t_e^k \cdot d^k = x^{k+1} - x^k \neq 0$  gilt, kann insbesondere keiner der Vektoren  $d^k$  verschwinden.

Im nächsten Schritt suchen wir nach Möglichkeiten, konjugierte Suchrichtungen explizit zu erzeugen. Der folgende Satz besagt, dass man konjugierte Richtungen zum Beispiel aus den Quasi-Newton-Verfahren der Broyden-Familie erhält.

**Satz 2.2.56.** Für  $\theta \geq 0$  werde Algorithmus 2.6 mit  $t^k = t_e^k$  und  $B^{k+1} = B_\theta^{k+1}$  auf  $q() = \frac{1}{2} x^T A x + b^T x$  mit  $A = A^T \succ 0$  angewendet, und für ein  $k \in \mathbb{N}$  seien die Iterierten  $x^0, \dots, x^k$  paarweise verschieden. Dann sind die Richtungen  $d^0, \dots, d^{k-1}$  paarweise konjugiert bezüglich  $A$  und sämtlich von null verschieden.

Bei Wahl exakter Schrittweiten minimieren die Quasi-Newton-Verfahren der Broyden-Familie konvex-quadratische Funktionen also in höchstens  $n$  Schritten. Für eine beliebige  $C^2$ -Funktion  $f$  lässt sich das dahingehend interpretieren, dass sie die lokale quadratische

Approximation an  $f$  in  $n$  Schritten minimieren, im Hinblick auf Übung 2.2.43 also einen Schritt des Newton-Verfahrens simulieren. Unter geeigneten Voraussetzungen und mit Neustarts nach jeweils  $n$  Schritten konvergieren sie daher „ $n$ -Schritt-quadratisch“.

## Konjugierte-Gradienten-Verfahren

Wir betrachten weiterhin das Abstiegsverfahren  $x^{k+1} = x^k + t_e^k d^k$  mit exakten Schrittweiten  $t_e^k$  und Abstiegsrichtungen erster Ordnung  $d^k$  für die konvex-quadratische Funktion

$$q(x) = \frac{1}{2} x^T A x + b^T x$$

mit  $A = A^T \succ 0$  und  $b \in \mathbb{R}^n$ . Gesucht sind Möglichkeiten, konjugierte Suchrichtungen ( $d^k$ ) zu erzeugen.

Die Grundidee wird im Folgenden sein, die Suchrichtungen  $d^k$  rekursiv zu wählen, nämlich als Kombination des aktuellen negativen Gradienten  $-\nabla q(x^k)$  und der letzten Suchrichtung  $d^{k-1}$  mit Hilfe eines noch zu bestimmenden „Gewichts“  $\alpha^k \in \mathbb{R}$  zu

$$d^k = -\nabla q(x^k) + \alpha_k \cdot d^{k-1}, k = 1, 2, \dots$$

Zu Beginn dieser Rekursion setzen wir  $d^0 = -\nabla q(x^0)$ . Das folgende Lemma wird zur Bestimmung der Werte  $\alpha_k$  wesentlich sein.

**Lemma 2.2.57.** Es seien  $d^0, \dots, d^{k-1}$  paarweise konjugiert bezüglich  $A$  und  $x^1, \dots, x^k$  schon generiert mit  $x^l \neq x^{l-1}$  für  $1 \leq l \leq k$ . Dann ist  $d^k$  genau dann konjugiert zu einem  $d^l$  mit  $0 \leq l \leq k-1$ , wenn folgendes erfüllt ist

$$\langle \nabla q(x^{l+1}) - \nabla q(x^l), d^k \rangle = 0$$

**Satz 2.2.58.** Unter den Voraussetzungen von Lemma 2.2.58 ist die Richtung  $d^k = -\nabla q(x^k) + \alpha_k \cdot d^{k-1}$  genau für

$$\alpha_k = \frac{\|\nabla q(x^k)\|_2^2}{\|\nabla q(x^{k-1})\|_2^2}$$

konjugiert zu den Vektoren  $d^0, \dots, d^{k-1}$ .

Satz 2.2.58 motiviert den Algorithmus 2.7, da er für  $f(x) = q(x) = \frac{1}{2} x^T A x + b^T x$  mit  $A = A^T \succ 0$  nach höchstens  $n$  Schritten den globalen Minimalpunkt liefert. Man benutzt dieses Verfahren zum Beispiel zur Lösung hochdimensionaler linearer Gleichungssysteme  $Ax = b$  durch den Kleinste-Quadrate-Ansatz, also per Minimierung von  $\|r(x)\|_2^2$  mit dem Residuum  $r(x) = Ax - b$ .

Wegen Rundungsfehlern bricht das Verfahren aber selten tatsächlich nach  $n$  Schritten ab, so dass auch seine Konvergenzgeschwindigkeit untersucht wurde. Es stellt sich heraus, dass sie von der Wurzel der Konditionszahl (also dem Quotienten aus größtem und kleinstem

Eigenwert) der Matrix  $A^T A$  abhängt. Es bietet sich daher an, das Gleichungssystem  $Ax = b$  zunächst so äquivalent umzuformen, dass diese Konditionszahl sinkt. Dies ist als Präkonditionierung bekannt.

---

**Algorithmus 2.7:** CG-Verfahren von Fletcher-Reeves

---

**Input :**  $C^1$ -Optimierungsproblem  $P$   
**Output :** Approximation  $\bar{x}$  eines kritischen Punkts von  $f$  (falls das Verfahren terminiert [25, Th. 5.7])

```

1 begin
2   Wähle einen Startpunkt  $x^0$ , eine Toleranz  $\varepsilon > 0$  und setze  $d^0 = -\nabla f(x^0)$  sowie  $k = 0$ .
3   while  $\|\nabla f(x^k)\| > \varepsilon$  do
4     Setze  $x^{k+1} = x^k + t_e^k d^k$ .
5     Setze  $d^{k+1} = -\nabla f(x^{k+1}) + (\|\nabla f(x^{k+1})\|_2^2 / \|\nabla f(x^k)\|_2^2) \cdot d^k$ .
6     Ersetze  $k$  durch  $k + 1$ .
7   end
8   Setze  $\bar{x} = x^k$ .
9 end

```

---

**Bemerkung 2.2.59.** Der Kleinste-Quadrate-Ansatz per CG-Verfahren zur Lösung linearer Gleichungssysteme  $Ax = b$  lässt sich auch auf überbestimmte Gleichungssysteme anwenden, die keine Lösung besitzen.

Entscheidend für die Einsetzbarkeit von Algorithmus 2.7 ist, dass für  $f = q$  nirgends explizit die Matrix  $A$  eingeht, aber trotzdem bezüglich  $A$  konjugierte Suchrichtungen erzeugt werden. Man kann das Verfahren also auch für beliebige  $C^1$ -Funktionen formulieren, wobei die Armijo-Schrittweite für gewöhnlich verwendet wird. Unter geeigneten Voraussetzungen erhält man wieder, dass  $n$  CG-Schritte einen Newton-Schritt simulieren, also „ $n$ -Schrittquadratische Konvergenz“.

## Trust-Region-Verfahren

Im Gegensatz zu klassischen Suchrichtungsverfahren wählen Trust-Region-Verfahren erst den Suchradius  $t$  und dann die Suchrichtung  $d$ . Dazu benutzt man in Iteration  $k$  des allgemeinen Abstiegsverfahrens aus Algorithmus 2.3 wie folgt ein quadratisches Modell für  $f$ .

Nach dem Satz von Taylor (Satz 2.1.30b) gilt für  $f \in C^2(\mathbb{R}^n, \mathbb{R})$

$$f(x^k + d) \approx f(x^k) + \langle \nabla f(x^k), d \rangle + \frac{1}{2} d^T D^2 f(x^k) d$$

Mit  $c^k := f(x^k)$ ,  $b^k = \nabla f(x^k)$  und einer symmetrischen Matrix  $A^k$  (zum Beispiel, aber nicht notwendigerweise,  $A^k = D^2 f(x^k)$ ) nennt man die Funktion

$$m^k(d) := c^k + \langle b^k, d \rangle + \frac{1}{2} d^T A^k d$$

ein lokales quadratisches Modell für  $f$  um  $x^k$ .

Man betrachtet daher  $m^k$  nur für  $\|d\|_2 \leq t^k$  und bestimmt man einen optimalen Punkt  $d^k$  des Trust-Region-Hilfsproblems

$$TR^k : \min_{d \in \mathbb{R}^n} m^k(d) \text{ s.t. } \|d\|_2 \leq t^k$$

Der folgende Quotient misst dabei die Güte der Approximation

$$r^k := \frac{f(x^k) - f(x^k + d^k)}{m^k(0) - m^k(d^k)}$$

- Ein Wert  $r^k < 0$  impliziert daher  $f(x^k + d^k) > f(x^k)$ , d. h.,  $x^{k+1} = x^k + d^k$  würde einen Anstieg im Zielfunktionswert liefern. Folglich ist die Trust Region zu groß, und ihr Radius  $t^k$  muss verkleinert werden.
- Liegt andererseits  $r^k$  nahe bei eins, dann beschreibt das lokale Modell die Funktion  $f$  sehr gut; man setzt  $x^{k+1} = x^k + d^k$  und vergrößert in der nächsten Iteration probeweise den Trust-Region-Radius  $t^k$ .
- Insbesondere für  $r^k \geq \frac{1}{4}$  wird  $t^k$  dort nicht verkleinert, und der Schritt wird angenommen, für  $r^k < 0$  wird  $t^k$  verkleinert, und der Schritt wird abgelehnt, und für  $r^k \in [0, \frac{1}{4})$  wird  $t^k$  verkleinert, und der Schritt wird dann abgelehnt, wenn  $r^k \leq \eta$  gilt.

Ein entscheidender Vorteil von Trust-Region-Verfahren gegenüber Variable-Metrik-Verfahren besteht allerdings darin, dass die Matrizen  $A^k$  nicht positiv definit zu sein brauchen. Insbesondere für  $A^k \equiv 0$  erhält man als „Trust-Region-Gradientenverfahren“ lediglich ein übliches Gradientenverfahren mit einer speziellen Schrittweitensteuerung. Von einem solchen Verfahren ist wegen Satz 2.2.23 keine schnelle Konvergenz zu erwarten.

Allerdings auch hier eine inexakte Lösung von  $TR^k$ , um globale Konvergenz zu gewährleisten. Eine Möglichkeit dafür besteht darin, die zulässige Menge von  $TR^k$  stark zu verkleinern und beispielsweise nur nichtnegative Vielfache der beim Verfahren gefundenen Suchrichtung zuzulassen:

$$TR_C^k : \min_{s \in \mathbb{R}} m^k \left( -\frac{s \cdot t^k}{\|\nabla f(x^k)\|_2} \cdot \nabla f(x^k) \right) \text{ s.t. } 0 \leq s \leq 1$$

Die Lösung zu diesem Problem ergibt sich  $d_C^k := -\frac{s^k t^k}{\|\nabla f(x^k)\|_2} \nabla f(x^k)$  mit

$$s^k := \begin{cases} 1, & \text{falls } Df(x^k) A^k \nabla f(x^k) \leq 0 \\ \min \left\{ \frac{\|\nabla f(x^k)\|_2^3}{t^k Df(x^k) A^k \nabla f(x^k)}, 1 \right\}, & \text{sonst} \end{cases}$$

---

**Algorithmus 2.8:** Trust-Region-Verfahren

---

**Input :**  $C^1$ -Optimierungsproblem  $P$ **Output :** Approximation  $\bar{x}$  eines kritischen Punkts von  $f$  (falls das Verfahren terminiert; Satz 2.2.63)

```
1 begin
2   Wähle einen Startpunkt  $x^0$ , eine Matrix  $A^0 = (A^0)^\top$ , eine Toleranz  $\varepsilon > 0$ , einen
   Maximalradius  $\check{t} > 0$ , einen Startradius  $t^0 \in (0, \check{t})$ , einen Parameter  $\eta \in [0, 1/4)$ 
   und setze  $k = 0$ .
3   while  $\|\nabla f(x^k)\|_2 > \varepsilon$  do
4     Berechne einen (inexakten) Optimalpunkt  $d^k$  von  $TR^k$  und setze


$$r^k = \frac{f(x^k) - f(x^k + d^k)}{m^k(0) - m^k(d^k)}.$$


5     if  $r^k < \frac{1}{4}$  then
6       Setze  $t^{k+1} = \frac{1}{4} \|d^k\|_2$ .
7     else
8       if  $r^k > \frac{3}{4}$  and  $\|d^k\|_2 = t^k$  then
9         Setze  $t^{k+1} = \min\{2t^k, \check{t}\}$ .
10      else
11        Setze  $t^{k+1} = t^k$ .
12      end
13    end
14    if  $r^k > \eta$  then
15      Setze  $x^{k+1} = x^k + d^k$ .
16    else
17      Setze  $x^{k+1} = x^k$ .
18    end
19    Wähle  $A^{k+1} = (A^{k+1})^\top$ .
20    Ersetze  $k$  durch  $k + 1$ .
21  end
22  Setze  $\bar{x} = x^k$ .
23 end
```

---

**Definition 2.2.60** (Cauchy-Punkt). Der Punkt  $x_C^{k+1} = x^k + d_C^k$  heißt **Cauchy-Punkt** zu  $x^k$  und  $t^k$ .

**Übung 2.2.61.** Der Vektor  $d^k = d_C^k$  erfüllt die folgende Ungleichung mit  $c = 0.5$

$$m^k(0) - m^k(d^k) \geq c \cdot \|\nabla f(x^k)\|_2 \cdot \min \left\{ t^k, \frac{\|\nabla f(x^k)\|_2}{\|A^k\|_2} \right\} \quad (2.24)$$

**Bemerkung 2.2.62.** Die exakte Lösung  $d_e^k$  von  $TR^k$  erfüllt wegen der Zulässigkeit von  $d_C^k$  für  $TR^k$  die Ungleichung  $m^k(d_e^k) \leq m^k(d_C^k)$  und damit nach Übung 2.2.61 ebenfalls (2.24) mit  $c = 0.5$ .

**Satz 2.2.63.** Die Menge  $f_{\leq}^{f(x^0)}$  sei beschränkt, die Funktion  $\nabla f$  sei Lipschitz-stetig auf  $\text{conv}(f_{\leq}^{f(x^0)})$ , die Folge  $(\|A^k\|_2)$  sei beschränkt, und die Folge  $(d^k)$  der inexakten Lösungen von  $TR^k$  erfülle (2.24) mit  $c > 0$ . Dann gilt in Algorithmus 2.8:

- a) Für  $\eta = 0$  ist  $\liminf_k \|\nabla f(x^k)\|_2 = 0$  (d.h.  $(x^k)$  besitzt einen Häufungspunkt  $x^*$  mit  $\nabla f(x^*) = 0$ ).
- b) Für  $\eta \in (0, 1)$  ist  $\lim_k \nabla f(x^k) = 0$  (d.h. alle Häufungspunkte von  $(x^k)$  sind kritisch).

Nach Übung 2.2.61, Bemerkung 2.2.62 und Satz 2.2.63 liefern sowohl die inexakten Lösungen  $d_C^k$  als auch die exakten Lösungen  $d_e^k$  von  $TR^k$  globale Konvergenz. Während die exakte Lösung  $d_e^k$  wie erwähnt schwer berechenbar sein kann, ist das Ausweichen auf die inexakte Lösung  $d_C^k$  selten ratsam, da die Matrix  $A^k$  lediglich die Länge von  $d_C^k$  beeinflusst und man so im Wesentlichen nach wie vor das Gradientenverfahren erhält.

### Dogleg-Methode

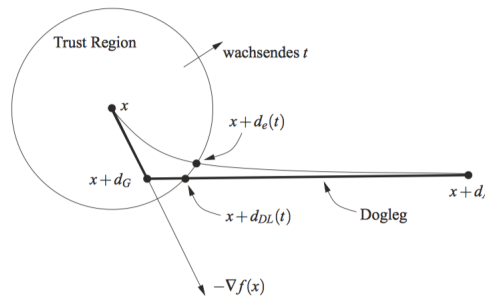
Sei  $A$  positiv definit und  $d_A := -A^{-1}\nabla f(x)$ . Die Dogleg-Methode approximiert diese Kurve durch einen Polygonzug von  $x$  nach  $x + d_A$  mit zwei Segmenten, wobei als Zwischenpunkt  $x + d_G$  mit dem exakten Minimalpunkt  $d_G$  von  $m$  entlang  $-\nabla f(x)$  gewählt wird, den man zu

$$d_G = -\frac{\|\nabla f(x)\|_2^2}{Df(x)A\nabla f(x)}\nabla f(x)$$

berechnet. Formal lautet der Polygonzug damit  $\{x + \hat{d}(s) \mid s \in [0, 2]\}$  mit

$$\hat{d}(s) = \begin{cases} s \cdot d_G, & 0 \leq s \leq 1 \\ d_G + (s - 1)(d_A - d_G), & 1 \leq s \leq 2 \end{cases}$$

**Abb. 2.10** Approximation der Kurve  $\{x + d_e(t) \mid t \geq 0\}$  per Dogleg



**Übung 2.2.64.** Es gilt für  $A = A^T \succ 0$ :

- a)  $\|\hat{d}(s)\|_2$  ist monoton wachsend in  $s$ .
- b)  $m(\hat{d}(s))$  ist monoton fallend in  $s$ .

Somit ergibt die Dogleg-Moethode die inexakte Lösung  $x + d_{DL}(t)$  mit

$$d_{DL} = \begin{cases} d_A, & \text{falls } \|d_A\|_2 < t \\ \|s \cdot d_G\|_2 = t & \text{falls } t < \|d_G\|_2 \\ \|d_G + (s - 1)(d_A - d_G)\|_2 = t, & \text{sonst} \end{cases}$$

### Minimierung auf einem zweidimensionalen Teilraum

Die inexakte Lösung von  $TR$  durch die Dogleg-Methode kann verbessert werden, indem man  $TR$  nicht auf den eindimensionalen Polygonzug einschränkt, sondern auf den zweidimensionalen Teilraum, der von  $d_G$  und  $d_A$  aufgespannt wird. In diesem Raum liegen insbesondere alle Punkte des Polygonzugs. Man erhält das Hilfsproblem

$$\min_{d \in \mathbb{R}^n} m(d) \text{ s.t. } \|d\|_2 \leq t, \quad d \in \text{bild}(\nabla f(x), A^{-1}\nabla f(x))$$

Ein Hauptvorteil dieses Ansatzes besteht darin, dass er sich im Gegensatz zur Dogleg-Methode sinnvoll auf indefinite Matrizen  $A$  erweitern lässt. Für Details sei auf [25] verwiesen.

## II. Restringierte Probleme

Wir betrachten im folgenden Problem

$$P : \min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } g_i(x) \leq 0, i \in I, h_j(x) = 0, j \in J$$

Mit von vektorwertigen Funktionen lässt sich die Menge  $M$  der für  $P$  zulässigen Punkte schreiben als

$$M = \{x \in \mathbb{R}^n \mid g(x) \leq 0, h(x) = 0\}$$

### Topologische Eigenschaften

In Übung 1.2.11 haben wir bereits gesehen, dass  $M$  unter der Stetigkeitsvoraussetzung an die Funktionen  $g$  und  $h$  eine abgeschlossene Menge ist.

**Definition 3.1.2** (Aktive-Index-Menge). Zu  $\bar{x} \in M$  heißt

$$I_0(\bar{x}) = \{i \in I \mid g_i(\bar{x}) = 0\}$$

Menge der aktiven Indizes oder auch **Aktive-Index-Menge**.

**Satz 3.1.3.** Für jedes  $\bar{x} \in M$  existiert eine Umgebung  $U$  von  $\bar{x}$  mit

$$U \cap M = U \cap \{x \in \mathbb{R}^n \mid g_i(x) \leq 0, i \in I_0(\bar{x}), h_j(x) = 0, j \in J\}$$

**Definition 3.1.4** (Zulässige Abstiegsrichtung). Gegeben sei das Problem

$$P : \min f(x) \text{ s.t. } x \in M$$

mit (nicht notwendigerweise in funktionaler Beschreibung vorliegender) zulässiger Menge  $M \subseteq \mathbb{R}^n$ . Dann heißt ein Vektor  $d \in \mathbb{R}^n$  **zulässige Abstiegsrichtung** für  $P$  in  $\bar{x} \in M$ , falls folgendes gilt

$$\exists \hat{t} > 0 \forall t \in (0, \hat{t}) : f(\bar{x} + td) < f(\bar{x}), \bar{x} + td \in M$$

**Übung 3.1.5.** Für das Problem  $P$  aus Definition 3.1.4 sei  $\bar{x}$  ein lokaler Minimalpunkt. Dann existiert keine zulässige Abstiegsrichtung für  $P$  in  $\bar{x}$ .

**Übung 3.1.6.** Gegeben sei das Problem

$$P : \min f(x) \text{ s.t. } g_i(x) \leq 0, i \in I, h_j(x) = 0, j \in J$$



Ein Vektor  $d \in \mathbb{R}^n$  ist genau dann zulässige Abstiegsrichtung für  $P$  in  $\bar{x} \in M$ , wenn folgendes gilt

$$\exists \hat{t} > 0 \forall t \in (0, \hat{t}) : f(\bar{x} + td) < f(\bar{x}), \bar{x} + td \in M$$

**Definition 3.1.7** (Äußerer Linearisierungskegel). Für  $\bar{x} \in \mathbb{R}^n$  heißt

$$L_{\leq}(\bar{x}, M) = \{d \in \mathbb{R}^n \mid \langle \nabla g_i(\bar{x}), d \rangle \leq 0, i \in I_0(\bar{x})\}$$

**äußerer Linearisierungskegel** an  $M$  in  $\bar{x}$ .

**Definition.** Eine Menge  $A \subseteq \mathbb{R}^n$  wird als Kegel bezeichnet, wenn

$$\forall a \in A, \lambda > 0 : \quad \forall \cdot a \in A$$

**Übung 3.1.8.** Am Punkt  $\bar{x} \in M$  seien die Funktionen  $g_i, i \in I_0(\bar{x})$ , differenzierbar, dann ist  $L_{\leq}(\bar{x}, M)$  ein konvexer Kegel.

Die funktionale Beschreibung einer zulässigen Menge kann so ungeschickt sein kann, dass ein äußerer Linearisierungskegel die lokale Struktur der Menge nicht notwendigerweise gut wiedergibt.

**Definition 3.1.11** (Innerer Linearisierungskegel). Für  $\bar{x} \in \mathbb{R}^n$  heißt

$$L_{<}(\bar{x}, M) = \{d \in \mathbb{R}^n \mid \langle \nabla g_i(\bar{x}), d \rangle < 0, i \in I_0(\bar{x})\}$$

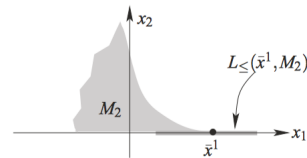
**innerer Linearisierungskegel** an  $M$  in  $\bar{x}$ .

**Definition 3.1.12** (Nichtdegenerierte funktionale Beschreibung einer Menge). Die funktionale Beschreibung von  $M$  heißt an  $\bar{x}$  **nichtdegeneriert**, wenn  $\text{cl } L_{<}(\bar{x}, M) = L_{\leq}(\bar{x}, M)$  gilt. Ansonsten heißt sie **degeneriert**. Diese Gleichheit ist auch als Cottle-Bedingung (Cottle constraint qualification) bekannt.

**Satz 3.1.15.** Die funktionale Beschreibung von  $M$  ist an  $\bar{x}$  genau dann nichtdegeneriert, wenn  $L_{<}(\bar{x}, M) \neq \emptyset$  gilt.

Es gibt zudem Fälle, in denen schon die Geometrie der zulässigen Menge so ungünstig ist, dass keine funktionale Beschreibung die gewünschte „gute“ Approximation erster Ordnung liefert.

**Abb. 3.4** Äußerer Linearisierungskegel in Beispiel 3.1.16



**Definition 3.1.16** (Innerer und äußerer Tangentialkegel). Es seien  $\bar{x} \in \mathbb{R}^n$  und  $M \subseteq \mathbb{R}^n$ . Eine Richtung  $\bar{d} \in \mathbb{R}^n$  liegt im

- a) **inneren Tangentialkegel**  $\Gamma(\bar{x}, M)$  an  $M$  in  $\bar{x}$ , falls ein  $\hat{t} > 0$  und eine Umgebung von  $\bar{d}$  existieren mit

$$\forall t \in (0, \hat{t}), d \in D : \quad \bar{x} + td \in M,$$

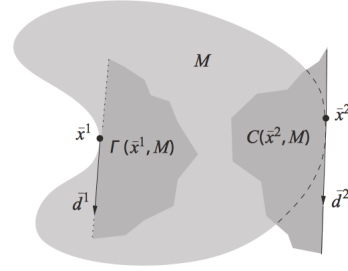
- b) **äußeren Tangentialkegel**  $C(\bar{x}, M)$  an  $M$  in  $\bar{x}$ , falls Folgen  $(t^k)$  und  $(d^k)$  existieren mit

$$t^k \searrow 0, d^k \rightarrow \bar{d}, \forall k \in \mathbb{N} : \quad \bar{x} + t^k d^k \in M$$

Der äußere Tangentialkegel ist im Gegensatz zum äußeren Linearisierungskegel nicht notwendigerweise konvex.

Abb. 3.5 zeigt Beispiele für innere und äußere Tangentialkegel sowie Vektoren  $\bar{d}^1 \notin \Gamma(\bar{x}^1, M)$  und  $\bar{d}^2 \in C(\bar{x}^2, M)$ . Würde man in Definition 3.1.17 keine variablen Richtungen zulassen, so resultierte dies hingegen in  $\bar{d}^1 \in \Gamma(\bar{x}^1, M)$  und  $\bar{d}^2 \notin C(\bar{x}^2, M)$ .

**Abb. 3.5** Innerer und äußerer Tangentialkegel



**Lemma 3.1.17.** Es seien  $\bar{x} \in \mathbb{R}^n$  und  $M \subseteq \mathbb{R}^n$ . Dann gilt:

- a)  $\Gamma(\bar{x}, M) \subseteq C(\bar{x}, M)$ .
- b)  $\Gamma(\bar{x}, M)^C = C(\bar{x}, M^c)$ .
- c)  $\Gamma(\bar{x}, M)$  ist ein offener und  $C(\bar{x}, M)$  ein abgeschlossener Kegel.

**Definition 3.1.18** (Nichtdegenerierte Geometrie einer Menge). Die Geometrie von  $M$  heißt an  $\bar{x}$  **nichtdegeneriert**, wenn  $\text{cl} \Gamma(\bar{x}, M) = C(\bar{x}, M)$  gilt. Ansonsten heißt sie **degeneriert**.

**Satz 3.1.19.** Für alle  $\bar{x} \in M$  gilt die Inklusions

$$L_{<}(\bar{x}, M) \subseteq \Gamma(\bar{x}, M) \subseteq C(\bar{x}, M) \subseteq L_{\leq}(\bar{x}, M).$$

**Korollar 3.1.20.** Die funktionale Beschreibung der Menge  $M$  sei an  $\bar{x}$  nichtdegeneriert. Dann ist auch die Geometrie von  $M$  an  $\bar{x}$  nichtdegeneriert.

# Optimalitätsbedingungen

## Stationarität

Da wir dabei nur die Geometrie von  $M$  benutzen wollen, definieren wir Stationarität mit Hilfe einer geometrischen Approximation erster Ordnung, nämlich des äußeren Tangentialkegels.

**Definition 3.2.1** (Stationärer Punkt - restringierter Fall). Die Funktion  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  sei an  $\bar{x} \in M$  differenzierbar. Dann heißt  $\bar{x}$  stationärer Punkt von  $P$ , falls  $\langle \nabla f(\bar{x}), d \rangle \geq 0$  für jede Richtung  $d \in C(\bar{x}, M)$  gilt.

**Satz 3.2.2.** Die Funktion  $f: \mathbb{R} \rightarrow \mathbb{R}$  sei an einem lokalen Minimalpunkt  $\bar{x}$  von  $P$  differenzierbar. Dann ist  $\bar{x}$  stationärer Punkt im Sinne von Definition 3.2.1.

Hier nutzen wir aus, dass differenzierbare Funktionen auch einseitige richtungsdifferenzierbar im Sinne von Hadamard sind:

$$\langle \nabla f(x), d \rangle = \lim_k \frac{f(x + t^k d^k) - f(x)}{t^k}$$

## Constraint Qualifications

Diese Beobachtungen führen zur Definition von zwei Regularitätsbedingungen (constraint qualifications).

**Definition 3.2.3** (Abadie- und Mangasarian-Fromowitz-Bedingung für  $J = \emptyset$ ). An  $\bar{x} \in M$  gilt

- a) die **Abadie-Bedingung** (AB) für  $J = \emptyset$ , falls folgendes erfüllt ist

$$C(\bar{x}, M) = L_{\leq}(\bar{x}, M)$$

- b) die **Mangasarian-Fromowitz-Bedingung** (MFB) für  $J = \emptyset$ , falls folgendes gilt

$$L_{<}(\bar{x}, M) \neq \emptyset$$

Nach Definition des inneren Linearisierungskegels ist die MFB an einem Punkt  $\bar{x}$  genau dann erfüllt, wenn eine Richtung  $d \in \mathbb{R}$  mit folgendem existiert

$$\langle \nabla g_i(\bar{x}), d \rangle < 0, \quad i \in I_0(\bar{x})$$

Die MFB kann an einem Punkt bloß deshalb verletzt sein kann, weil die dort geometrisch nichtdegenerierte zulässige Menge degeneriert funktional beschrieben ist.

**Korollar 3.2.4.** An einem lokalen Minimalpunkt  $\bar{x}$  von  $P$  seien  $f$  und die Funktionen  $g_i$ ,  $i \in I_0(\bar{x})$  differenzierbar.

a) Dann ist das folgende System mit keinem  $d \in \mathbb{R}^n$  lösbar:

$$\langle \nabla f(\bar{x}), d \rangle < 0, \quad \langle \nabla g_i(\bar{x}), d \rangle < 0 \quad i \in I_0(\bar{x})$$

b) Falls an  $\bar{x}$  die AB gilt, dann ist sogar das folgende System mit keinem  $d \in \mathbb{R}^n$  lösbar:

$$\langle \nabla f(\bar{x}), d \rangle < 0, \quad \langle \nabla g_i(\bar{x}), d \rangle \leq 0 \quad i \in I_0(\bar{x})$$