08_Grundlagen_Optimierung

Martin Reißel

27. Juni 2022

Inhaltsverzeichnis

1 Grundlagen der Optimierung

1.1 Überblick

In diesem Abschnitt betrachten wir verschiedene Methoden aus der Analysis zur Behandlung von Optimierungsproblemen.

1.2 Grundlagen

Wir betrachten eine Zielfunktion $f : \mathbb{R}^d \to \mathbb{R}$. Minimiert man

- f über ganz $\mathbb{R}^d$, so liegt ein *nicht restringiertes* Problem vor
- f über $\emptyset \neq X \subset \mathbb{R}^d$, so handelt es sich um ein *restringiertes* Problem

Ist $f(x) \geq c > -\infty$, so existiert $f_* = \inf_X f$, aber nicht notwendig ein x_* mit $f(x_*) = \inf_X f = f_*$

Beispiel:

- für $f(x) = e^{-x^2}$, $x \in \mathbb{R}$, ist
$$\inf_{x \in \mathbb{R}} f(x) = 0$$
aber es gibt kein $x_* \in \mathbb{R}$ mit $f(x_*) = 0$.

Ist f stetig, $X \subset \mathbb{R}^d$ kompakt, dann hat das restringierte Problem (mindestens) eine Lösung, d.h.

$$\operatorname{argmin}_{x \in X} f(x) \neq \emptyset$$

1.3 Nicht restringierte Probleme

Im nicht restringierten Fall erhält man bei (ausreichend oft) differenzierbarer Zielfunktion f notwendige und hinreichende Bedingungen für (lokale) Minimalstellen.

Ist $x_* \in \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$, dann gelten die notwendigen Bedingungen:

- ist $f \in C^1(\mathbb{R}^d) \Rightarrow f'(x_*) = 0$

- ist $f \in C^2(\mathbb{R}^d) \Rightarrow f'(x_*) = 0$ und $f''(x_*)$ ist positiv semidefinit

Als hinreichende Bedingungen erhalten wir für $f \in C^2(\mathbb{R}^d)$:

- ist $f'(\tilde{x}) = 0$ und $f''(\tilde{x})$ positiv definit oder $f''(x)$ in einer offenen Umgebung U von $\tilde{x}$ positiv semidefinit, dann gilt
$$\tilde{x} \in \operatorname{argmin}_{x \in U} f(x),$$
d.h. $\tilde{x}$ ist lokales Minimum

1.4 Restringierte Probleme, Lagrange-Funktion, KKT-Bedingungen

Das Äquivalent zu den obigen Ergebnissen bei der restringierten Optimierung sind die *Karush-Kuhn-Tucker-Bedingungen*.

Wir betrachten das Optimierungsproblem

$$\min_{x \in X} f(x), \quad g(x) \leq 0, \quad h(x) = 0,$$

mit $\emptyset \neq X \subset \mathbb{R}^d$ und

$$f : X \to \mathbb{R}, \quad g : X \to \mathbb{R}^m, \quad h : X \to \mathbb{R}^p.$$

Wir nehmen an, dass $f, g, h \in C^1(X)$ und definieren die zugehörige *Lagrange-Funktion* L durch

$$L(x, \lambda, \mu) = f(x) + \lambda^T g(x) + \mu^T h(x).$$

x_*, λ_*, μ_* heißt KKT-Punkt, falls

$$\begin{aligned} \partial_x L(x_*, \lambda_*, \mu_*) &= 0, \\ g(x_*) &\leq 0 \\ h(x_*) &= 0, \\ \lambda_* &\geq 0, \\ \lambda_{*,i} g_i(x_*) &= 0 \quad \forall i \end{aligned}$$

Ist x_* eine Lösung des Optimierungsproblems und erfüllt gewisse Regularitätsbedingungen (Constraint-Qualifications), dann existieren $\lambda_* \geq 0, \mu_*$, so dass x_*, λ_*, μ_* ein KKT-Punkt ist. Wir haben es hier also mit notwendige Bedingungen zu tun.

Sind X, f, g_i konvex und h linear affin, dann vereinfacht sich das ganze noch:

- als Constraint-Qualification muss nur die *Slater-Bedingung* gelten, d.h.
$$\exists \tilde{x} \in X \quad \text{mit} \quad g(\tilde{x}) < 0, \quad h(\tilde{x}) = 0$$
- KKT ist auch hinreichend, d.h.x_* ist lokales (und damit auch globales) Minimum, ohne weitere Annahmen

Beispiel:

- wir betrachten das Problem
$$\min_{x \in \mathbb{R}^2,\ x_1^2 + x_2^2 \leq 1,\ x_1 = x_2} x_1$$
- mit $X = \mathbb{R}^2$,
$$f(x) = x_1, \quad g(x) = x_1^2 + x_2^2 - 1, \quad h(x) = x_1 - x_2$$
erhalten wir die Standardform
$$\min_{x \in X} f(x), \quad g(x) \leq 0, \quad h(x) = 0$$
- X, f, g sind konvex, h ist linear affin und die Slater-Bedingung ist ebenfalls erfüllt

- als Lagrange-Funktion erhalten wir

$$L(x,\lambda,\mu) = x_1 + \lambda(x_1^2 + x_2^2 - 1) + \mu(x_1 - x_2)$$

  und somit

$$\partial_x L(x,\lambda,\mu) = \begin{pmatrix} 1 + 2\lambda x_1 + \mu \\ 2\lambda x_2 - \mu \end{pmatrix} = 0,$$
$$g(x) = x_1^2 + x_2^2 - 1 \le 0,$$
$$h(x) = x_1 - x_2 = 0,$$
$$\lambda g(x) = \lambda\left(x_1^2 + x_2^2 - 1\right) = 0$$

- subtrahiert man im Gradienten die zweite von der ersten Gleichung, so folgt

$$1 + 2\lambda(x_1 - x_2) + 2\mu = 0$$

  und wegen $x_1 - x_2 = 0$ schließlich

$$\mu = -\frac{1}{2}$$

- eingesetzt in die zweite Komponente des Gradienten erhalten wir

$$\lambda x_2 = \frac{\mu}{2} = -\frac{1}{4}$$

- aus $\lambda \ge 0$ folgt daraus $\lambda > 0$ und

$$x_2 < 0$$

  und wegen $\lambda\left(x_1^2 + x_2^2 - 1\right) = 0$ somit

$$x_1^2 + x_2^2 = 1$$

- außerdem soll noch die Gleichheitsbedingung

$$x_1 = x_2$$

  gelten

- insgesamt erhalten wir damit

$$x_2 = -\frac{1}{\sqrt{2}}$$

  und somit ist

$$x_* = -\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

  wegen der Konvexität und der Slater-Bedingung die (eindeutige) Lösung des Optimierungsproblems

Im nicht-konvexen Fall gelten für $f, g, h \in C^2$ folgende hinreichende Bedingungen:

- ist x_*, λ_*, μ_* ein KKT-Punkt und gilt

$$s^T \partial_x^2 L(x_*, \lambda_*, \mu_*)\, s \ge 0$$

  für alle $s \ne 0$ mit

$$\begin{pmatrix} \partial_x g_i(x_*) \\ \partial_x h_j(x_*) \end{pmatrix}^T s = 0$$

  mit i s.d. $\lambda_{*,i} > 0$, dann ist x_* ein lokales Minimum

1.5 Dualität

Wir betrachten das primale Problem
$$\min_{x\in X} f(x), \quad g(x) \le 0, \quad h(x) = 0,$$
die zugehörige Lagrange-Funktion
$$L(x,\lambda,\mu) = f(x) + \lambda^T g(x) + \mu^T h(x)$$
und definieren damit die *duale Funktion*
$$q(\lambda,\mu) = \inf_{x\in X} L(x,\lambda,\mu).$$

Für q untersuchen wir das duale Problem
$$\max_{\lambda\ge 0,\mu} q(\lambda,\mu)$$
über dem *wesentlichen Zulässigkeitsbereich*
$$\mathrm{dom}_q = \left\{(\lambda,\mu) \;\middle|\; \lambda \ge 0,\ q(\lambda,\mu) > -\infty\right\}$$

Das duale Problem hat folgende Eigenschaften:

- dom_q ist immer konvex
- $-q$ ist immer konvex
- ist
$$R_p = \left\{x \;\middle|\; g(x) \le 0, h(x) = 0\right\}$$
der zulässige Bereich des primalen Problems, dann gilt
$$\sup_{\mathrm{dom}_q} q(\lambda,\mu) \le \inf_{R_p} f(x)$$
und somit gilt für alle Lösungen λ_*, μ_* von $\sup_{\mathrm{dom}_q} q(\lambda,\mu)$ und x_* von $\inf_{R_p} f(x)$
$$q(\lambda_*,\mu_*) \le f(x_*).$$

Die Differenz
$$f(x_*) - q(\lambda_*,\mu_*) \ge 0$$
bezeichnet man als *Dualitäts-Lücke*. Gilt
$$f(x_*) - q(\lambda_*,\mu_*) = 0$$
so spricht man von starker Dualität:

- starke Dualität gilt nicht immer
- sind f, g konvex, h linear affin und die Slater-Bedingung erfüllt, dann gilt starke Dualität

Praktischer Einsatz:

- ist nur der Wert $f_* = f(x_*)$ (und nicht x_*) interessant, dann kann man statt des primalen das duale Problem lösen, das immer konvex ist
- ist $\bar{x}$ eine Näherung von x_* und gilt starke Dualität, so kann man mit Hilfe geeignet gewählter $\bar{\lambda}, \bar{\mu}$ über
$$f(\bar{x}) - q(\bar{\lambda},\bar{\mu})$$
einen Fehlerindikator für $\bar{x}$ berechnen

Beispiel:

- wir betrachten
$$\min_{x\in\mathbb{R}^d} c^T x, \quad Ax - b \le 0, \quad A \in \mathbb{R}^{m\times d}$$

- als Lagrange-Funktion erhalten wir mit $\lambda \in \mathbb{R}^m$, $\lambda \geq 0$

$$\begin{aligned} L(x,\lambda) &= c^T x + \lambda^T (Ax - b) \\ &= (c^T + \lambda^T A)x - \lambda^T b \end{aligned}$$

  bzw. mit $y = c + A^T \lambda$

$$L(x,\lambda) = y^T x - \lambda^T b$$

- jetzt berechnen wir die duale Funktion

$$q(\lambda) = \inf_{x \in \mathbb{R}^d} L(x,\lambda)$$

- L ist linear affin in x mit

$$\partial_x L(x,\lambda) = y^T$$

- ist $y \neq 0$ so erhalten wir mit $x = \nu y$

$$L(x,\lambda) = \nu \|y\|_2^2 - \lambda^T b \xrightarrow{\nu \to -\infty} -\infty$$

- für $y = 0$ (also $\partial_x L(x,\lambda) = 0$) gilt

$$L(x,\lambda) = -\lambda^T b = \inf_{x \in \mathbb{R}^d} L(x,\lambda)$$

- somit ist

$$q(\lambda) = \begin{cases} -\lambda^T b & \text{für } c + A^T \lambda = 0 \\ -\infty & \text{sonst} \end{cases}$$

  und das duale Problem hat die Form

$$\max_{\lambda \in \mathbb{R}^m} (-\lambda^T b), \quad \lambda \geq 0, \quad A^T \lambda = -c$$

Beispiel:

- wir leiten jetzt die duale Form des Optimierungsproblems bei Support-Vector Classifiern her, die wir oben im Zusammenhang mit dem Kernel-Trick benutzt haben
- das primale Problem lautet

$$\min_{v \neq 0, v_0, \xi} \left(\frac{1}{2} \|v\|_2^2 + C \sum_{i=1}^n \xi_i \right)$$

  mit

$$y_i(v^T x_i + v_0) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n.$$

- setzen wir

$$\begin{aligned} w &= (v, v_0, \xi) \in X = \mathbb{R}^{m+1+n} \\ f(w) &= \frac{1}{2} \|v\|_2^2 + C \sum_{i=1}^n \xi_i \\ g_i(w) &= -\Big(y_i \big(\sum_{j=1}^m v_j x_{ij} + v_0 \big) - 1 + \xi_i \Big), \quad i = 1, \ldots, n \\ g_i(w) &= -\xi_i, \quad i = n+1, \ldots, 2n \end{aligned}$$

  so erhalten wir die Standardform

$$\min_{w \in \mathbb{R}^{m+1+n}} f(w), \quad g(w) \leq 0$$

- mit $\lambda = (\alpha, \beta)$, $\alpha, \beta \geq 0$, folgt für die Lagrange-Funktion

$$\begin{aligned} L(w,\lambda) =& L(v, v_0, \xi, \alpha, \beta) \\ =& \frac{1}{2}\|v\|_2^2 + C\sum_{i=1}^n \xi_i \\ & - \sum_{i=1}^n \alpha_i \Big(y_i \big(\sum_{j=1}^m v_j x_{ij} + v_0 \big) - 1 + \xi_i \Big) \\ & - \sum_{i=1}^n \beta_i \xi_i \\ =& \frac{1}{2}\|v\|_2^2 - \sum_{i=1}^n \alpha_i y_i \sum_{j=1}^m v_j x_{ij} \\ & - v_0 \sum_{i=1}^n \alpha_i y_i \\ & + \sum_{i=1}^n (C - \alpha_i - \beta_i)\xi_i \\ & + \sum_{i=1}^n \alpha_i \end{aligned}$$

  bzw. für die duale Funktion

$$q(\alpha, \beta) = \inf_{(v, v_0, \xi) \in \mathbb{R}^{m+1+n}} L(v, v_0, \xi, \alpha, \beta)$$

- L ist quadratisch in v und linear affin in v_0 und ξ (und somit konvex und differenzierbar)
- wie im vorherigen Beispiel folgt, dass $q(\alpha, \beta)$ nur dann größer $-\infty$ sein kann, wenn

$$\begin{aligned} 0 &= \partial_{v_0} L(v, v_0, \xi, \alpha, \beta) = -\sum_{i=1}^n \alpha_i y_i, \\ 0 &= \partial_{\xi_i} L(v, v_0, \xi, \alpha, \beta) = C - \alpha_i - \beta_i, \quad i = 1, \ldots, n, \end{aligned}$$

  gilt, insgesamt also

$$\alpha_i, \beta_i \geq 0, \quad \sum_{i=1}^n \alpha_i y_i = 0, \quad \alpha_i + \beta_i = C$$

- unter diesen Einschränkungen vereinfacht sich L zu

$$L(v, v_0, \xi, \alpha, \beta) = \frac{1}{2}\|v\|_2^2 - \sum_{i=1}^n \alpha_i y_i \sum_{j=1}^m v_j x_{ij} + \sum_{i=1}^n \alpha_i,$$

  ist also insbesondere unabhängig von v_0, ξ, β und wird bei gegebenem α minimal an $\hat{v}$ mit

$$0 = \partial_{v_k} L(\hat{v}, v_0, \xi, \alpha, \beta) = \hat{v}_k - \sum_{i=1}^n \alpha_i y_i x_{ik}, \quad k = 1, \ldots, m,$$

  also

$$\hat{v}_k = \sum_{i=1}^n \alpha_i y_i x_{ik}.$$

- eingesetzt in L erhalten wir damit

$$
\begin{aligned}
q(\alpha,\beta) &= L(\hat{v}, v_0, \xi, \alpha, \beta) \\
&= \frac{1}{2}\|\hat{v}\|_2^2 - \sum_{i=1}^{n} \alpha_i y_i \sum_{j=1}^{m} \hat{v}_j x_{ij} + \sum_{i=1}^{n} \alpha_i \\
&= \frac{1}{2}\sum_{j=1}^{m} \hat{v}_j^2 - \sum_{j=1}^{m} \hat{v}_j \sum_{i=1}^{n} \alpha_i y_i x_{ij} + \sum_{i=1}^{n} \alpha_i \\
&= \frac{1}{2}\sum_{j=1}^{m} \hat{v}_j^2 - \sum_{j=1}^{m} \hat{v}_j^2 + \sum_{i=1}^{n} \alpha_i \\
&= -\frac{1}{2}\sum_{j=1}^{m} \hat{v}_j^2 + \sum_{i=1}^{n} \alpha_i \\
&= -\frac{1}{2}\sum_{j=1}^{m} \Big(\sum_{i=1}^{n} \alpha_i y_i x_{ij} \Big)^2 + \sum_{i=1}^{n} \alpha_i \\
&= -\frac{1}{2}\sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i y_i \Big(\sum_{j=1}^{m} x_{ij} x_{kj} \Big) y_k \alpha_k + \sum_{i=1}^{n} \alpha_i
\end{aligned}
$$

  bzw.

$$
q(\alpha,\beta) = -\frac{1}{2}\alpha^T Q \alpha + e^T \alpha
$$

  mit

$$
Q = \big(y_i x_i^T x_k y_k\big)_{i,k=1,\ldots,n}, \quad e = (1,\ldots,1)^T
$$

- insgesamt erhalten wir für die duale Funktion q für $\alpha, \beta \geq 0$

$$
q(\alpha,\beta) = \begin{cases} -\frac{1}{2}\alpha^T Q\alpha + e^T\alpha & \text{für} \quad y^T\alpha = 0, \quad \alpha + \beta = Ce \\ -\infty & \text{sonst} \end{cases}
$$

- als duales Problem haben wir damit

$$
\max_{\alpha\geq 0, \beta \geq 0} q(\alpha,\beta) = \min_{\alpha \geq 0} \Big(\frac{1}{2}\alpha^T Q\alpha - e^T \alpha \Big)
$$

  unter den Nebenbedingungen

$$
y^T\alpha = 0, \quad \alpha_i + \beta_i = C \quad \forall i
$$

- da $\beta_i \geq 0$ nur in die Nebenbedingung eingeht, kann diese auf $0 \leq \alpha_i \leq C$ geändert werden
- somit erhalten wir die finale Form des dualen Problems, wie wir sie bei den Support-Vector Classifiern bereits benutzt haben:

$$
\min_{\alpha\geq 0} \Big(\frac{1}{2}\alpha^T Q\alpha - e^T\alpha \Big)
$$

  mit

$$
Q = \big(y_i x_i^T x_j y_j\big)_{i,j=1,\ldots,n}, \quad y^T\alpha = 0, \quad 0 \leq \alpha \leq Ce.
$$

1.6 Zusammenfassung

Für nicht-restringierte bzw, restringierte Optimierungsprobleme haben wir die Werkzeuge aus der Analysis betrachtet, u.a. Lagrange-Funktion, KKT-Bedingungen und duale Probleme.