AGML, Sommersemester 2020 Vorlesung 9 (5 Seiten)

Gradient Descent

Thomas Kesselheim Letzte Aktualisierung: 22. Mai 2020

In der letzten Vorlesung haben wir in Form von Hard- und dem Soft-SVM-Problem bereits zwei konvexe Optimierungsprobleme kennengelernt. Im Maschinellen Lernen gibt es eine Vielzahl weiterer derartiger Probleme. Heute werden wir diskutieren, mit welchen algorithmischen Ansätzen man sie lösen kann.

Allgemein ist ein konvexes Optimierungsproblem wie folgt definiert. Wir müssen eine konvexe Funktion $f\colon S \to \mathbb{R}$ minimieren, wobei $S \subseteq \mathbb{R}^n$ die (konvexe) Menge aller zulässigen Lösungen darstellt. Wir beschränken uns auf den Fall, dass $S = \mathbb{R}^n$. Das heißt, es gibt keine Nebenbedingungen.

Zunächst beschränken wir uns auf differenzierbare Funktionen f. Später werden wir jedoch unsere Ergebnisse verallgemeinern, dass sie auch mit nicht-differenzierbaren Funktionen anwendbar sind.

1 Gradienten

Betrachten wir zunächst eine differenzierbare Funktion f. Folglich hat sie einen Gradienten ∇f, der der Vektor aller partiellen Ableitungen ist; $(\nabla f(\mathbf{u}))_i = \frac{\partial f}{\partial u_i}(\mathbf{u})$. Konvexität von f ist nun äquivalent dazu, dass für alle $\mathbf{u}, \mathbf{v}$

$$f(\mathbf{u}) \geq f(\mathbf{v}) + \langle \nabla f(\mathbf{v}), (\mathbf{u} - \mathbf{v}) \rangle \ . \tag{1}$$

Zum Verständnis dieser Ungleichung ist es hilfreich zu verstehen, dass

$$\mathbf{u} \mapsto f(\mathbf{v}) + \langle \nabla f(\mathbf{v}), (\mathbf{u} - \mathbf{v}) \rangle$$

die lineare Approximation von f durch die Tangentialhyperebene an der Stelle $\mathbf{v}$ ist. Das heißt, eine konvexe Funktion muss jeweils oberhalb der Tangentialhyperebene liegen.

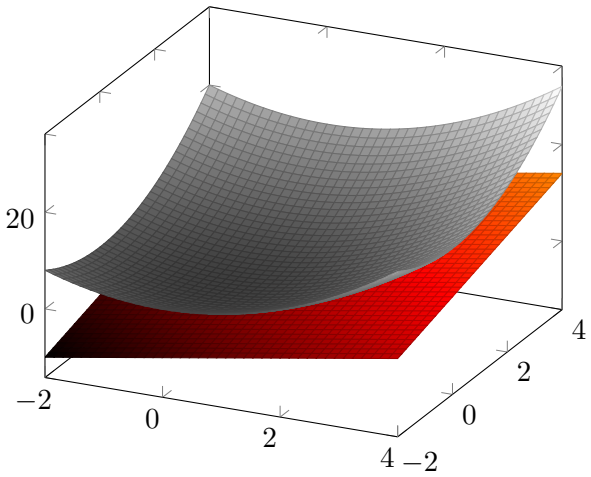

Abbildung 1: $f\colon \mathbb{R}^2 \to \mathbb{R}$ mit $f(x_1, x_2) = x_1^2 + x_2^2$ und die Tangentialebene an $(1, 1)$.

Beispiel 9.1. *In Abbildung 1 ist die Funktion* $f\colon \mathbb{R}^2 \to \mathbb{R}$ *mit* $f(x_1, x_2) = x_1^2 + x_2^2$ *und die Tangentialebene an* f *an* $(1, 1)$ *dargestellt. Der Gradient ist* $\nabla f(x_1, x_2) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix}$*, entsprechend ist die Tangentialebene an* $(1, 1)$ *gegeben durch*

$$\mathbf{u} \mapsto 2 + 2(u_1 - 1) + 2(u_2 - 1) \ .$$

Abbildung 2 zeigt eine andere Darstellung einer Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ als Relief-Plot. Hier sehen wir Höhenlinien der Funktion eingetragen, also Mengen von Punkten, an denen die Funktion denselben Wert hat. Der Gradient im Punkt $\mathbf{x}$ steht immer senkrecht zur Höhenlinie im Punkt $\mathbf{x}$ der Funktion. Er zeigt in die Richtung des stärksten Anstiegs

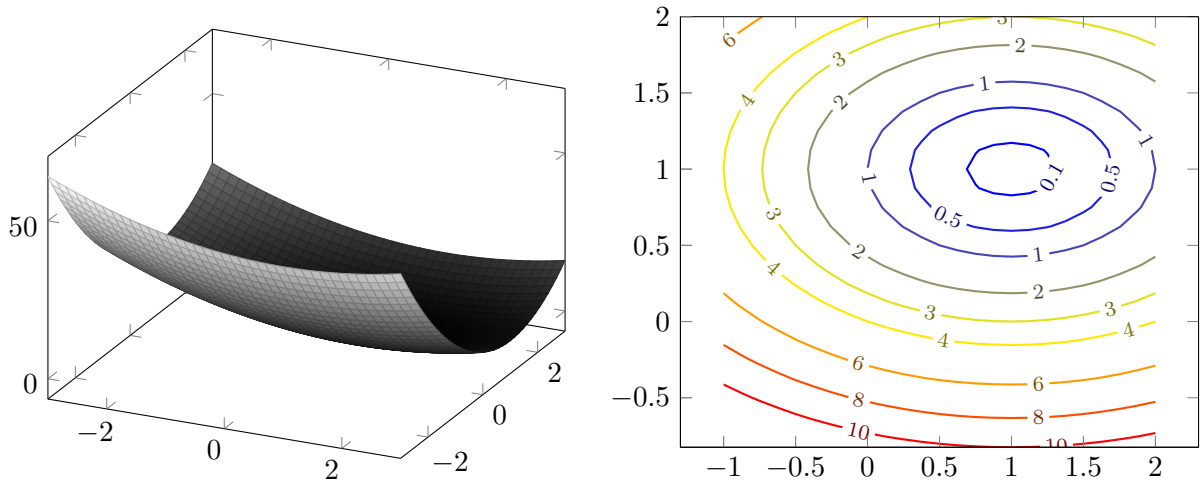

Abbildung 2: Links ein 3D-Plot, rechts ein Relief-Plot mit Höhenlinien der konvexen Funktion $f\colon \mathbb{R}^2 \to \mathbb{R}$ definiert durch $f(x_1, x_2) = (x_1 - 1)^2 + 3(x_2 - 1)^2$.

2 Gradient Descent

Der Algorithmus *Gradient Descent* berechnet eine Folge von Lösungen $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)}$. Wir beginnen mit $\mathbf{w}^{(1)} = 0$. Die Lösung $\mathbf{w}^{(t+1)}$ ergibt sich jeweils aus einer leichten Verbesserung von $\mathbf{w}^{(t)}$.

Betrachten wir hierfür den Gradienten $\mathbf{g}^{(t)} := \nabla f(\mathbf{w}^{(t)})$ von f an der Stelle $\mathbf{w}^{(t)}$. Weil der Gradient in die Richtung des stärksten Anstiegs zeigt, müssen wir uns in die entgegengesetzte Richtung, also $-\mathbf{g}^{(t)}$ bewegen, denn dies ist die Richtung des stärksten Abfalls. Dies führt zur Regel

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{g}^{(t)} \ .$$

Dabei ist η (ausgesprochen: *eta*) ein Parameter des Algorithmus. Wenn wir η zu klein wählen, machen wir keine guten Fortschritte. Wenn wir η zu groß wählen, schießen wir möglicherweise über das Ziel hinaus.

Nach einer festen Anzahl von Iterationen T geben wir die beste gesehene Lösung zurück.[1]

[1] Alternative Formulierungen des Algorithmus geben einen Durchschnitt über alle Lösungen oder die letzte erreichte Lösung zurück.

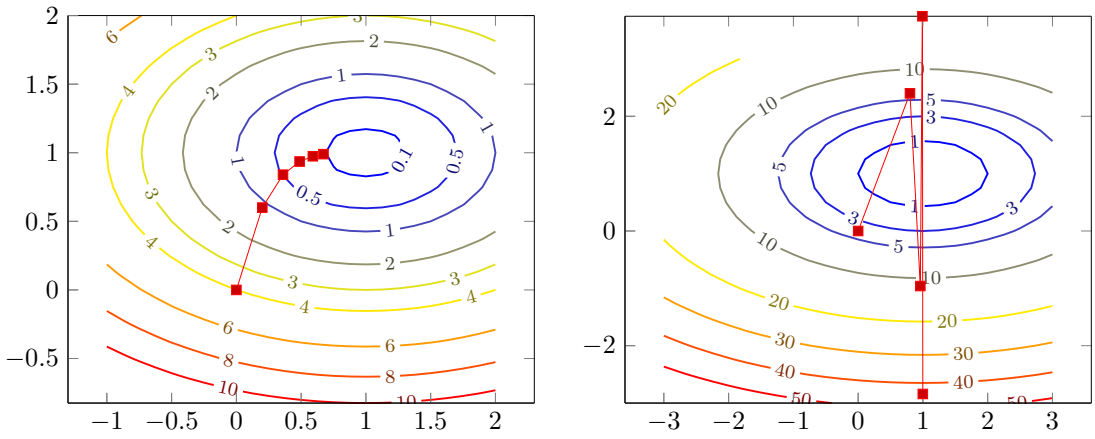

Abbildung 3: Beispiel von Gradient Descent auf $f(x_1, x_2) = (x_1 - 1)^2 + 3(x_2 - 1)^2$, links mit $\eta = 0.3$ mit Konvergenz, rechts mit $\eta = 0.4$ ohne Konvergenz.

3 Analyse von Gradient Descent

Wir können nun zeigen, dass der Algorithmus sich tatsächlich einer optimale Lösung annähert.

Satz 9.2. *Gilt* $\|\mathbf{g}^{(t)}\| \leq \rho$ *für alle* t*, dann gilt für alle* $\mathbf{w}^* \in \mathbb{R}^n$ *mit* $\|\mathbf{w}^*\| \leq B$

$$\min_t f(\mathbf{w}^{(t)}) \leq f(\mathbf{w}^*) + \frac{B^2}{2\eta T} + \frac{\eta \rho^2}{2} \ .$$

Insbesondere gilt für $\eta = \frac{B}{\rho\sqrt{T}}$

$$\min_t f(\mathbf{w}^{(t)}) \leq f(\mathbf{w}^*) + \frac{B\rho}{\sqrt{T}} \ .$$

Insbesondere können wir natürlich $\mathbf{w}^*$ als die optimale Lösung wählen und erhalten damit einen additiven Fehler von höchstens $\frac{B\rho}{\sqrt{T}}$ unter den genannten Bedingungen. Wichtig ist an dieser Stelle, dass der Fehler immer kleiner wird je größer T, also die Anzahl der Iterationen, wird. Die Bedeutungen von B und ρ werden wir später noch diskutieren.

Beweis. Den besten gesehenen Funktionswert können wir abschätzen durch den durchschnittlich gesehenen Funktionswert

$$\min_t \left(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*) \right) \leq \frac{1}{T} \sum_{t=1}^{T} \left(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*) \right) \ . \tag{2}$$

Nun folgt der einzige Schritt, in dem wir Konvexität nutzen. Gemäß dieser gilt für alle t

$$f(\mathbf{w}^*) \geq f(\mathbf{w}^{(t)}) + \left\langle \mathbf{g}^{(t)}, \mathbf{w}^* - \mathbf{w}^{(t)} \right\rangle \ . \tag{3}$$

Also gilt

$$f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*) \leq \left\langle \mathbf{g}^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle \ .$$

Dieses Skalarprodukt drücken wir nun in einer Summe von Vektornormen aus. Es gilt nämlich für alle $\mathbf{u}, \mathbf{v}$, dass

$$\langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle = \|\mathbf{u} + \mathbf{v}\|^2 \ ,$$

aber auch

$$\langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v}\rangle = \langle \mathbf{u},\mathbf{u}\rangle + \langle \mathbf{v},\mathbf{u}\rangle + \langle \mathbf{u},\mathbf{v}\rangle + \langle \mathbf{v},\mathbf{v}\rangle = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2\langle \mathbf{u},\mathbf{v}\rangle \ .$$

Zusammengenommen also

$$\langle \mathbf{u},\mathbf{v}\rangle = \frac{1}{2}\left(\|\mathbf{u}+\mathbf{v}\|^2 - \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2\right) \ .$$

Mittels der Gleichung können wir nun schreiben

$$\left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, -\eta\mathbf{g}^{(t)}\right\rangle = \frac{1}{2}\left(\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\mathbf{g}^{(t)}\|^2 - \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|-\eta\mathbf{g}^{(t)}\|^2\right) \ .$$

Wir können nun $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta\mathbf{g}^{(t)}$ einsetzen. Zusätzlich teilen wir die Gleichung durch $-\eta$. Somit ergibt sich

$$\begin{aligned}\left\langle \mathbf{g}^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^*\right\rangle &= -\frac{1}{\eta}\left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, -\eta\mathbf{g}^{(t)}\right\rangle \\ &= -\frac{1}{2\eta}\left(\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|-\eta\mathbf{g}^{(t)}\|^2\right) \\ &= \frac{1}{2\eta}\left(\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2\right) + \frac{\eta}{2}\|\mathbf{g}^{(t)}\|^2 \ .\end{aligned}$$

Als Teleskopsumme ergibt sich damit für (2) zusammen mit (3)

$$\begin{aligned}\sum_{t=1}^{T}\left(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)\right) &\le \frac{1}{2\eta}\sum_{t=1}^{T}\left(\|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}^{(t)}\|^2 \\ &= \frac{1}{2\eta}\left(\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}^{(t)}\|^2 \ .\end{aligned}$$

Mit $\mathbf{w}^{(1)} = 0$ und $\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 \ge 0$ können wir also abschätzen

$$\sum_{t=1}^{T}\left(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)\right) \le \frac{1}{2\eta}\|\mathbf{w}^*\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{g}^{(t)}\|^2 \ .$$

Insgesamt erhalten wir damit

$$\min_t\left(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)\right) \le \frac{B^2}{2\eta T} + \frac{\eta\rho^2}{2} \ .$$

Und mit $\eta = \frac{B}{\rho\sqrt{T}}$ gilt nun $\min_t\left(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)\right) \le \frac{B\rho}{\sqrt{T}}$. □

4 Nicht-Differenzierbare Funktionen

Ist eine Funktion nicht differenzierbar, so gibt es nicht an jeder Stelle $\mathbf{v}$ einen Gradienten $\nabla f(\mathbf{v})$. Somit ist auch die Tangentialhyperebene nicht (eindeutig) definiert. In Abbildung 4 ist die Betragsfunktion dargestellt. An der Stelle 0 ist sie nicht differenzierbar. Es gibt nun eine Vielzahl von Tangenten, die wir an dieser Stelle anlegen können. Die Abbildung zeigt zwei Beispiele.

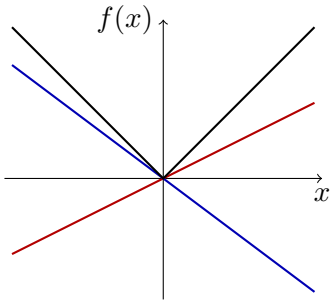

Abbildung 4: Die Betragsfunktion mit zwei möglichen Tangenten an der Stelle 0. Die Funktion liegt oberhalb von allen diesen Tangenten.

Wir erinnern uns, das Konvexität bei differenzierbaren Funktionen f äquivalent dazu ist, dass für alle $\mathbf{u}, \mathbf{v}$

$$f(\mathbf{u}) \geq f(\mathbf{v}) + \langle \nabla f(\mathbf{v}), (\mathbf{u} - \mathbf{v}) \rangle \ .$$

Diese Ungleichung haben wir für die Analyse von Gradient Descent genutzt.

Für allgemeine, nicht notwendigerweise differenzierbare Funktionen gibt es glücklicherweise folgende Verallgemeinerung: Eine Funktion f ist konvex, wenn es für alle $\mathbf{v}$ ein $\mathbf{g}$ gibt, sodass für alle $\mathbf{u}$ gilt

$$f(\mathbf{u}) \geq f(\mathbf{v}) + \langle \mathbf{g}, (\mathbf{u} - \mathbf{v}) \rangle \ . \tag{4}$$

Wenn f in $\mathbf{v}$ differenzierbar ist, dann ist tatsächlich $\mathbf{g} = \nabla f(\mathbf{v})$ die einzige Wahl, die diese Ungleichung erfüllt. Ist f in $\mathbf{v}$ nicht differenzierbar, gibt es möglicherweise mehrere Möglichkeiten, $\mathbf{g}$ zu wählen.

Definition 9.3. *Für eine Funktion $f \colon S \to \mathbb{R}$ und $\mathbf{v} \in S$ nennen wir*

$$\partial f(\mathbf{v}) = \{\mathbf{g} \mid f(\mathbf{u}) \geq f(\mathbf{v}) + \langle \mathbf{g}, (\mathbf{u} - \mathbf{v}) \rangle \text{ für alle } \mathbf{u} \in S\}$$

das Subdifferenzial *von f in $\mathbf{v}$. Die Elemente von $\partial f(\mathbf{v})$ heißen* Subgradienten.

Eine Funktion f ist also genau dann konvex, wenn $\partial f(\mathbf{v}) \neq \emptyset$ für alle $\mathbf{v}$. Dies liegt daran, dass Wahlen für $\mathbf{g}$ in Ungleichung (4) genau den Elementen aus $\partial f(\mathbf{v})$ entsprechen.

5 Subgradient Descent

Der Algorithmus Subgradient Descent funktioniert genauso wie Gradient Descent. Der einzige Unterschied ist die Wahl von $\mathbf{g}^{(t)}$. Galt bisher die Regel, dass $\mathbf{g}^{(t)}$ auf $\nabla f(\mathbf{w}^{(t)})$ gesetzt wurde, ist nun $\mathbf{g}^{(t)} \in \partial f(\mathbf{w}^{(t)})$ beliebig. Das heißt, dass wir anstatt des Gradienten nun einen beliebigen Subgradienten verwenden. Für differenzierbare Funktionen ändert sich damit nichts.

Satz 9.2 und sein Beweis gelten weiterhin. Lediglich in Ungleichung (3) haben wir die Definition von $\mathbf{g}^{(t)}$ genutzt. Diese Ungleichung entspricht jedoch genau der Definition des Subgradienten.

Referenzen

- Understanding Machine Learning, Kapitel 14.1–14.2