AGML, Sommersemester 2020 Vorlesung 10 (6 Seiten)

Stochastic Gradient Descent

Thomas Kesselheim Letzte Aktualisierung: 26. Mai 2020

Wir betrachten heute wie der Gradient-Descent-Algorithmus auf dem Soft-SVM-Problem abläuft. Wir werden in diesem Zusammenhang eine Verallgemeinerung des Algorithmus namens *Stochastic Gradient Descent* kennenlernen, die schnellere Laufzeiten ermöglicht.

1 Soft-SVM: Wiederholung und neue Notation

Wir erinnern uns, dass uns beim Soft-SVM-Problem eine Menge S von Datenpunkten mit Labels $\mathbf{z}_1 = (\mathbf{x}_1, y_1), \dots, \mathbf{z}_m = (\mathbf{x}_m, y_m)$ gegeben ist, wobei $\mathbf{x}_i \in \mathbb{R}^d$ und $y_i \in \{-1, +1\}$ für alle i. Das Ziel ist es nun $\mathbf{w} \in \mathbb{R}^d$ und $u \in \mathbb{R}$ zu finden, so dass

$$\lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - u)\}$$

minimiert wird, wobei λ ein Parameter ist. Um die Notation einfach zu halten, fordern wir im Folgenden $u = 0$. Dies ist mehr oder weniger ohne Beschränkung der Allgemeinheit, wenn wir u als die $d+1$-te Komponente von $\mathbf{w}$ interpretieren und an alle $\mathbf{x}_i$ als letzte Komponente 1 anfügen. Zu einem anderen Zeitpunkt werden wir diese Anspekte noch genauer diskutieren.

Führen wir an dieser Stelle etwas Notation ein. Definiere nun

$$\ell^{\text{hinge}}(h_{\mathbf{w}}, \mathbf{z}_i) = \max\{0, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\} \ ,$$

das ausdrückt, „wie falsch“ die Hypothese $h_{\mathbf{w}}$ auf dem i-ten Datenpunkt $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ ist. Diese Funktion nennt sich *Hinge Loss*. Der Name bezieht sich darauf, dass der Funktionsgraph aussieht wie ein Türscharnier (siehe Abbildung 1). Der durchschnittliche Loss auf S ist nun

$$L_S^{\text{hinge}}(h_{\mathbf{w}}) = \frac{1}{m} \sum_{i=1}^{m} \ell^{\text{hinge}}(h_{\mathbf{w}}, \mathbf{z}_i) \ .$$

Wir müssen also $\mathbf{w} \in \mathbb{R}^d$ finden, sodass $f(\mathbf{w}) := R(\mathbf{w}) + L_S^{\text{hinge}}(h_{\mathbf{w}})$ minimiert wird, wobei $R(\mathbf{w}) = \lambda \|\mathbf{w}\|^2$. Auf die Bedeutung von $R(\mathbf{w})$ werden wir in einer späteren Vorlesung eingehen.

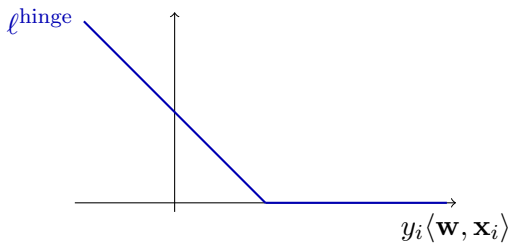

Abbildung 1: Die Hinge-Loss-Funktion.

2 Gradient Descent für Soft-SVM

Diese Funktion f ist konvex. Wir können also Gradient Descent nutzen, um sie zu minimieren. Genauer gesagt müssen wir Subgradient Descent nutzen, denn sie ist nicht überall differenzierbar.

Betrachten wir der Einfachheit halber eine Stelle $\mathbf{w}$, an der sie differenzierbar ist. Der Gradient ist der Vektor aller partiellen Ableitungen. Die partielle Ableitung nach w_j können wir mittels der üblichen Rechenregeln berechnen

$$\frac{\partial}{\partial w_j} f(\mathbf{w}) = \frac{\partial}{\partial w_j} R(\mathbf{w}) + \frac{\partial}{\partial w_j} L_S^{\text{hinge}}(h_{\mathbf{w}}) = \frac{\partial}{\partial w_j} R(\mathbf{w}) + \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial w_j} \ell^{\text{hinge}}(h_{\mathbf{w}}, \mathbf{z}_i) \ . \tag{1}$$

Weiterhin gelten

$$\frac{\partial}{\partial w_j} R(\mathbf{w}) = 2\lambda w_j \quad \text{und} \quad \frac{\partial}{\partial w_j} \ell^{\text{hinge}}(h_{\mathbf{w}}, \mathbf{z}_i) = \begin{cases} -y_i x_{i,j} & \text{falls } 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0 \\ 0 & \text{falls } 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle < 0 \end{cases}$$

Also gilt insgesamt

$$\nabla f(\mathbf{w}) = 2\lambda \mathbf{w} - \frac{1}{m} \sum_{i : 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0} y_i \mathbf{x}_i \ .$$

Wenn wir dies also in die Iterationsvorschrift von Gradient Descent $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)})$ einsetzen, ergibt sich

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \left(2\lambda \mathbf{w}^{(t)} - \frac{1}{m} \sum_{i : 1 - y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle > 0} y_i \mathbf{x}_i \right) = (1 - 2\eta\lambda) \mathbf{w}^{(t)} + \frac{\eta}{m} \sum_{i : 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0} y_i \mathbf{x}_i \ .$$

Der Algorithmus ist so also überraschend einfach. Hinsichtlich der Laufzeit einer einzelnen Iteration stellen wir fest, dass diese durch die Berechnung des Gradienten dominiert wird. Pro Dimension benötigen wir lineare Zeit in der Anzahl Samples m, insgesamt also $\Theta(dm)$. Das Problem hierbei ist, dass m typischerweise sehr groß sein sollte, denn die Stärke des Maschinellen Lernens liegt genau darin, aus der großen Menge an verfügbaren Daten Schlüsse zu ziehen.

3 Stochastic (Sub-) Gradient Descent

Die aufwändige Berechnung des Gradienten können wir wie folgt umgehen. Wie wir in Gleichung (1) sehen, ergibt sich die partielle Ableitung der Funktion f aus dem Durchschnitt der partiellen Ableitungen der Loss-Funktionen der einzelnen Datenpunkte. Diese Durchschnitt ersetzen wir nun durch ein Zufallsexperiment: Wir ziehen einen einzelnen Datenpunkt $\mathbf{z}_i$ und betrachten nur die partielle Ableitung, die sich für diesen einzelnen Punkt ergibt. Im Erwartungswert ergibt sich damit genau die gewünschte partielle Ableitung und damit auch Richtung für Gradient Descent.

Allgemeiner funktioniert der Algorithmus *Stochastic Gradient Descent* für eine beliebige konvexe Funktion f wie folgt. Wir beginnen wieder mit $\mathbf{w}^{(1)} = 0$. In Schritt t bestimmen wir $\mathbf{w}^{(t+1)}$ aus $\mathbf{w}^{(t)}$ wie folgt.

- Ziehe einen Vektor $\mathbf{g}^{(t)}$ aus irgendeiner Wahrscheinlichkeitsverteilung, sodass $\mathbf{E}\left[\mathbf{g}^{(t)} \mid \mathbf{w}^{(t)}\right] \in \partial f(\mathbf{w}^{(t)})$.[1]
- Setze $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{g}^{(t)}$.

[1] Diese Notation bedeutet, dass der *bedingte* Erwartungswert betrachtet wird. Der Vektor $\mathbf{w}^{(t)}$ wird also festgehalten und nun wird ein weiteres Zufallsexperiment durchgeführt, das von $\mathbf{w}^{(t)}$ abhängt.

4 Stochastic Subgradient Descent angewendet auf Soft-SVM

Im Fall von Soft-SVM hatten wir ja für Gradient Descent

$$\mathbf{g}^{(t)} = \nabla R(\mathbf{w}^{(t)}) + \frac{1}{m}\sum_{i=1}^{m} \nabla \ell^{\text{hinge}}(h_{\mathbf{w}^{(t)}}, \mathbf{z}_i)$$

gesetzt. Nun ziehen wir in jedem Schritt t ein I_t unabhängig, identisch verteilt aus $\{1, \dots, m\}$ und setzen

$$\mathbf{g}^{(t)} = \nabla R(\mathbf{w}^{(t)}) + \nabla \ell^{\text{hinge}}(h_{\mathbf{w}^{(t)}}, \mathbf{z}_{I_t}) = 2\lambda \mathbf{w}^{(t)} + \begin{cases} -y_{I_t}\mathbf{x}_{I_t} & \text{falls } 1 - y_{I_t}\langle \mathbf{w}^{(t)}, \mathbf{x}_{I_t}\rangle > 0 \\ 0 & \text{sonst} \end{cases} . \quad (2)$$

Anders formuliert erhalten wir

$$\mathbf{w}^{(t+1)} = \begin{cases} (1-\eta\lambda)\mathbf{w}^{(t)} + \eta y_i \mathbf{x}_i & \text{falls } 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle > 0 \\ (1-\eta\lambda)\mathbf{w}^{(t)} & \text{sonst} \end{cases} .$$

Nun gilt

$$\mathbf{E}\left[\mathbf{g}^{(t)} \;\middle|\; \mathbf{w}^{(t)}\right] = \sum_{i=1}^{m} \mathbf{Pr}\left[I_t = i\right]\left(\nabla R(\mathbf{w}^{(t)}) + \nabla \ell^{\text{hinge}}(h_{\mathbf{w}^{(t)}}, \mathbf{z}_i)\right) = \nabla R(\mathbf{w}^{(t)}) + \frac{1}{m}\sum_{i=1}^{m} \nabla \ell^{\text{hinge}}(h_{\mathbf{w}^{(t)}}, \mathbf{z}_i) .$$

Der bedingte Erwartungswert von $\mathbf{g}^{(t)}$ ist somit also genau der Gradient, den Gradient Descent nutzen würde.

5 Analyse von Stochastic (Sub-) Gradient Descent

Die allgemeine Formulierung von Stochastic (Sub-) Gradient Descent fordert nur $\mathbf{E}\left[\mathbf{g}^{(t)} \;\middle|\; \mathbf{w}^{(t)}\right] \in \partial f(\mathbf{w}^{(t)})$. Eine Möglichkeit wäre es also auch, den Vektor $\mathbf{g}^{(t)}$ deterministisch zu bestimmen als einen Subgradienten von f. Genau dies macht der Algorithmus Gradient Descent bzw. Subgradient Descent. Stochastic (Sub-) Gradient Descent ist also eine Verallgemeinerung. Trotzdem können wir genau dieselbe Garantie herleiten.

Satz 10.1. *Gilt $\|\mathbf{g}^{(t)}\| \leq \rho$ für alle t mit Wahrscheinlichkeit 1, dann gilt für alle $\mathbf{w}^* \in \mathbb{R}^n$ mit $\|\mathbf{w}^*\| \leq B$*

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

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

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

Wir erhalten also im Wesentlichen die gleiche Garantie wie bei Gradient Descent mit dem Unterschied, dass sie nur im Erwartungswert gilt. Das folgende Lemma fasst die wesentliche Änderung im Argument zusammen.

Lemma 10.2. *Bei Stochastic (Sub-) Gradient Descent gilt für alle t*

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

Beweis. Betrachten wir Schritt t und halten wir die Zufallsereignisse, die bis hier geschehen sind fest. Mathematisch formuliert, betrachten wir also den bedingten Wahrscheinlichkeitsraum für ein festes $\mathbf{w}^{(t)}$. Sei nun $\bar{\mathbf{g}} = \mathbf{E}\left[\mathbf{g}^{(t)} \mid \mathbf{w}^{(t)}\right]$. Gemäß unserer Annahme gilt $\bar{\mathbf{g}} \in \partial f(\mathbf{w}^{(t)})$. Das heißt insbesondere

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

und somit

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

Nun ist $\bar{g}_i = \mathbf{E}\left[g_i^{(t)} \;\middle|\; \mathbf{w}^{(t)}\right]$, also gilt wegen Linearität des Erwartungswerts

$$\begin{aligned}
\left\langle \bar{\mathbf{g}}, \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle &= \sum_{i=1}^{n} \bar{g}_i (\mathbf{w}^{(t)} - \mathbf{w}^*)_i \\
&= \sum_{i=1}^{n} \mathbf{E}\left[g_i^{(t)} \;\middle|\; \mathbf{w}^{(t)}\right] (\mathbf{w}^{(t)} - \mathbf{w}^*)_i \\
&= \mathbf{E}\left[\sum_{i=1}^{n} g_i^{(t)} (\mathbf{w}^{(t)} - \mathbf{w}^*)_i \;\middle|\; \mathbf{w}^{(t)}\right] \\
&= \mathbf{E}\left[\left\langle \mathbf{g}^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle \;\middle|\; \mathbf{w}^{(t)}\right] \quad .
\end{aligned}$$

Damit gilt für jedes $\mathbf{w}^{(t)}$, egal wie wir es erreicht haben

$$f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*) \leq \mathbf{E}\left[\left\langle \mathbf{g}^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle \;\middle|\; \mathbf{w}^{(t)}\right] \quad .$$

Um nun die Rechnung unkompliziert formal korrekt zu halten, nehmen wir an, dass $\mathbf{w}^{(t)}$ nur endlich viele Werte $\mathbf{v}_1, \ldots, \mathbf{v}_k$ und $\mathbf{g}^{(t)}$ nur endlich viele Werte $\mathbf{g}_1, \ldots, \mathbf{g}_\ell$ annehmen kann. Dann gilt für den unbedingten Erwartungswert

$$\begin{aligned}
\mathbf{E}\left[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)\right] &= \sum_{i=1}^{k} \mathbf{Pr}\left[\mathbf{w}^{(t)} = \mathbf{v}_i\right] (f(\mathbf{v}_i) - f(\mathbf{w}^*)) \\
&\leq \sum_{i=1}^{k} \mathbf{Pr}\left[\mathbf{w}^{(t)} = \mathbf{v}_i\right] \mathbf{E}\left[\left\langle \mathbf{g}^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle \;\middle|\; \mathbf{w}^{(t)} = \mathbf{v}_i\right] \\
&= \sum_{i=1}^{k} \mathbf{Pr}\left[\mathbf{w}^{(t)} = \mathbf{v}_j\right] \sum_{j=1}^{\ell} \mathbf{Pr}\left[\mathbf{g}^{(t)} = \mathbf{g}_j \;\middle|\; \mathbf{w}^{(t)} = \mathbf{v}_i\right] \langle \mathbf{g}_j, \mathbf{v}_i - \mathbf{w}^* \rangle \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} \mathbf{Pr}\left[\mathbf{w}^{(t)} = \mathbf{v}_j, \mathbf{g}^{(t)} = \mathbf{g}_j\right] \langle \mathbf{g}_j, \mathbf{v}_i - \mathbf{w}^* \rangle \\
&= \mathbf{E}\left[\left\langle \mathbf{g}^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle\right] \quad .
\end{aligned}$$

Diese Rechnung gilt auch allgemeiner. Dafür müssten wir allerdings den bedingten Erwartungswert formaler definieren, was über die Inhalte der Vorlesung hinausgeht. □

Nun können wir den Algorithmus im Wesentlichen wie Gradient Descent analysieren. Wir müssen lediglich des öfteren Gebrauch davon machen, dass der Erwartungswert linear ist.

Beweis von Satz 10.1. In der Analyse von Gradient Descent haben wir gezeigt, dass für all $\mathbf{u}, \mathbf{v}$ gilt

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

Diese Gleichung haben wir wie folgt genutzt, um $\left\langle \mathbf{g}^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^* \right\rangle$ umzuschreiben. Dabei ist es unerheblich, wie $\mathbf{g}^{(t)}$ definiert ist. Wir nutzen lediglich $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{g}^{(t)}$.

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

Ebenfalls erhalten wir über die Teleskopsumme und $\mathbf{w}^{(1)} = 0$ und $\|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2 \geq 0$ wieder

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

Nun können wir diese Gleichung mit Lemma 10.2 kombinieren. Aufgrund der Linearität des Erwartungswertes erhalten wir

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

Weil $\|\mathbf{w}^*\|^2 \leq B^2$ und $\mathbf{E} \left[\|\mathbf{g}^{(t)}\|^2 \right] \leq \rho^2$ gemäß Annahme, folgt der Satz. □

6 Norm des Subgradienten

Die Garantie in Satz 10.1 hängt von ρ ab, wobei wir fordern, dass $\|\mathbf{g}^{(t)}\| \leq \rho$ für alle t mit Wahrscheinlichkeit 1. Wie können wir diese Werte im Fall von Soft-SVM beschränken?

Betrachten wir Gleichung (2), können wir $\mathbf{g}^{(t)}$ schreiben als $\mathbf{g}^{(t)} = 2\lambda \mathbf{w}^{(t)} + \mathbf{v}^{(t)}$, wobei

$$\mathbf{v}^{(t)} = \begin{cases} -y_{I_t} \mathbf{x}_{I_t} & \text{falls } 1 - y_{I_t} \langle \mathbf{w}, \mathbf{x}_{I_t} \rangle > 0 \\ 0 & \text{sonst} \end{cases} \ .$$

Wir können also mittels der Dreiecksungleichung abschätzen

$$\|\mathbf{g}^{(t)}\| \leq 2\lambda\|\mathbf{w}^{(t)}\| + \|\mathbf{v}^{(t)}\| \leq 2\lambda\|\mathbf{w}^{(t)}\| + \max_i \|\mathbf{x}_i\| \ .$$

Entscheidend ist also, wie groß $\|\mathbf{w}^{(t)}\|$ werden kann. Dies ergibt sich aus dem bisherigen Verlauf des Algorithmus. Hierfür können wir $\mathbf{g}^{(t-1)}, \ldots, \mathbf{g}^{(1)}$ einsetzen und erhalten

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \eta\left(2\lambda\mathbf{w}^{(t-1)} + \mathbf{v}^{(t-1)}\right) = (1-2\eta\lambda)\,\mathbf{w}^{(t-1)} - \eta\mathbf{v}^{(t-1)} = \ldots = \sum_{i=1}^{t-1}(1-2\eta\lambda)^{t-1-i}\,\eta\mathbf{v}^{(i)} \ .$$

Nun erhalten wir mittels Dreiecksungleichung und geometrischer Summenformel

$$\|\mathbf{w}^{(t)}\| \leq \sum_{i=1}^{t-1}(1-2\eta\lambda)^{t-1-i}\,\eta\|\mathbf{v}^{(i)}\| \leq \sum_{i=0}^{\infty}(1-2\eta\lambda)^{i}\,\eta\max_i\|\mathbf{x}_i\| = \frac{1}{2\eta\lambda}\eta\max_i\|\mathbf{x}_i\| = \frac{1}{2\lambda}\max_i\|\mathbf{x}_i\| \ .$$

In die obige Schranke auf $\|\mathbf{g}^{(t)}\|$ eingesetzt, bekommen wir also

$$\|\mathbf{g}^{(t)}\| \leq 2\max_i\|\mathbf{x}_i\| \ .$$

Referenzen

- Understanding Machine Learning, Kapitel 14.3 und 14.5