12_Subgradient_Descent

Martin Reißel

27. Juni 2022

Inhaltsverzeichnis

1 Subgradient Descent

1.1 Überblick

Bisher haben wir immer $f \in C^1(\mathbb{R}^d)$ vorausgesetzt. In der Praxis trifft das nicht immer zu (Lasso, RELU-MLP).

Ist f konvex, dann existiert in jedem Punkt der Subgradient. Wir benutzen jetzt bei Gradient Descent statt des (eventuell nicht existierenden) Gradienten einen Subgradienten.

Die fehlende Glattheit von f wird sich negativ auf die Konvergenzgeschwindigkeit auswirken.

1.2 Grundlagen

Definition:

- ist $f : \mathbb{R}^d \supset \mathrm{dom}(f) \to \mathbb{R}$, dann ist $g \in \mathbb{R}^d$ *Subgradient* von f in $x \in \mathrm{dom}(f)$, falls
$$f(y) \geq f(x) + g^T(y-x) \quad \forall y \in \mathrm{dom}(f)$$
- die Menge aller Subgradienten von f an der Stelle x bezeichnen wir mit
$$\partial f(x) = \left\{ g \mid g \text{ ist Subgradient von } f \text{ in } x \right\}$$

Beispiel: Ist $f(x) = |x|$, dann ist $\partial f(0) = [-1, 1]$, denn
$$|y| = f(y) \geq f(0) + g(y-0) = gy \quad \forall y \in \mathbb{R}$$
genau dann wenn
$$g \in [-1, 1].$$

Lemma: Ist f differenzierbar in x, dann gilt $\partial f(x) \subset \{f'(x)\}$, also $\partial f(x) = \{f'(x)\}$ oder $\partial f(x) = \varnothing$.

Die Ungleichung $f(y) \geq f(x) + g^T(y-x)$ sieht aus wie die Konvexitätsbedingung für C^1-Funktionen, nur dass $f'(x)$ durch g ersetzt wurde. Dies ist kein Zufall.

Lemma: $f : \mathbb{R}^d \supset \mathrm{dom}(f) \to \mathbb{R}$ ist konvex genau dann, wenn $\mathrm{dom}(f)$ konvex ist und

$$\partial f(x) \neq \emptyset \quad \forall x \in \mathrm{dom}(f).$$

Beweis:

"$\Rightarrow$"

- ist f konvex, so ist $\mathrm{epi}(f)$ konvex
- $\partial f(x) \neq \emptyset$ folgt aus dem Existenzsatz für Stützhyperebenen für konvexe Mengen angewandt auf $\mathrm{epi}(f)$

"$\Leftarrow$"

- nach Voraussetzung ist $\mathrm{dom}(f)$ konvex und $\partial f(x) \neq \emptyset$, d.h. es existiert $g(x) \in \partial f(x)$ mit

$$f(y) \geq f(x) + g(x)(y-x) \quad \forall x, y \in \mathrm{dom}(f)$$

- für $t \in [0,1]$ setzen wir

$$z = (1-t)x + ty$$

- dann ist $z \in \mathrm{dom}(f)$ und

$$\begin{aligned} f(x) &\geq f(z) + g(z)(x-z) \\ f(y) &\geq f(z) + g(z)(y-z) \end{aligned}$$

- multiplizieren wir die erste Ungleichung mit $1-t$, die zweite mit t und addieren die Ergebnisse, dann erhalten wir

$$\begin{aligned} (1-t)f(x) + tf(y) &\geq (1-t)\big(f(z) + g(z)(x-z)\big) \\ &\quad + t\big(f(z) + g(z)(y-z)\big) \\ &= f(z) + g(z)\big(\underbrace{(1-t)x + ty}_{z} - z\big) \\ &= f(z) \\ &= f\big((1-t)x + ty\big) \end{aligned}$$

□

Existenz von Subgradienten und Konvexität sind also äquivalent.

Einige weitere Eigenschaften des Gradienten lassen sich auch auf Subgradienten übertragen. Für Lipschitz-Stetigkeit bei C^1-Funktionen f hatten wir $\|f'(x)\| \leq L_f$. Dies gilt analog auch für Subgradienten.

Lemma: $f : \mathbb{R}^d \supset \mathrm{dom}(f) \to \mathbb{R}$ konvex, $\mathrm{dom}(f)$ offen. Dann ist

$$\|f(x) - f(y)\| \leq L_f \|x-y\| \quad \forall x, y \in \mathrm{dom}(f)$$

äquivalent zu

$$\|g\| \leq L_f \quad \forall g \in \partial f(x) \quad \forall x \in \mathrm{dom}(f).$$

Beweis:

"$\Rightarrow$"

- für $x \in \mathrm{dom}(f)$, $g \in \partial f(x)$, $\varepsilon > 0$ definieren wir

$$y = \begin{cases} x + \varepsilon \frac{g}{\|g\|} & g \neq 0 \\ x & g = 0 \end{cases}$$

- damit folgt

$$y - x = \begin{cases} \varepsilon \frac{g}{\|g\|} & g \neq 0 \\ 0 & g = 0 \end{cases}$$

bzw.

$$\|y - x\| = \begin{cases} \varepsilon & g \neq 0 \\ 0 & g = 0 \end{cases} \leq \varepsilon$$

- wegen $\mathrm{dom}(f)$ offen ist für $\varepsilon > 0$ hinreichend klein auch $y \in \mathrm{dom}(f)$
- außerdem gilt

$$g^T(y - x) = \begin{cases} \varepsilon \frac{g^T g}{\|g\|} & g \neq 0 \\ 0 & g = 0 \end{cases} = \varepsilon \|g\|$$

- wegen $g \in \partial f(x)$ ist dann

$$f(y) \geq f(x) + g^T(y - x) = f(x) + \varepsilon \|g\|$$

- da f Lipschitz-stetig ist folgt

$$L_f \varepsilon \geq L_f \|y - x\| \geq f(y) - f(x) \geq \varepsilon \|g\|,$$

also

$$\|g\| \leq L_f$$

"$\Leftarrow$"

- für $x, y \in \mathrm{dom}(f)$ und $g \in \partial f(x)$ gilt

$$f(y) \geq f(x) + g^T(y - x)$$

also

$$f(x) - f(y) \leq g^T(x - y)$$

- mit Cauchy-Schwartz und $\|g\| \leq L_f$ folgt

$$f(x) - f(y) \leq \|g\| \|x - y\| \leq L_f \|x - y\|$$

- Vertauschen von x und y liefert schließlich

$$|f(x) - f(y)| \leq L_f \|x - y\| \quad \forall x, y \in \mathrm{dom}(f)$$

□

Satz: $f : \mathbb{R}^d \supset \mathrm{dom}(f) \to \mathbb{R}$, $x \in \mathrm{dom}(f)$, dann gilt

$$0 \in \partial f(x) \quad \Leftrightarrow \quad x \text{ ist globales Minimum von } f.$$

Beweis:

"$\Rightarrow$"

- ist $0 \in \partial f(x)$, dann folgt aus der Definition des Subgradienten

$$f(y) \geq f(x) + 0^T(y - x) = f(x) \quad \forall y \in \mathrm{dom}(f)$$

"$\Leftarrow$"

- ist $x \in \mathrm{dom}(f)$ globales Minimum, dann gilt

$$f(y) \geq f(x) = f(x) + 0^T(y - x) \quad \forall y \in \mathrm{dom}(f)$$

□

Bemerkung:

- die Bedingung $0 \in \partial f(x)$ ist stärker als $f'(x) = 0$ bei C^1-Funktionen
- $f'(x) = 0$ ist nur eine *notwendige Bedingung* für *lokale* Minima, d.h. aus $f'(x) = 0$ alleine kann man i.A. nicht folgern, dass x ein lokales oder gar globales Minimum ist
- dagegen garantiert $0 \in \partial f(x)$ immer, dass x globales Minimum ist

1.3 Vorüberlegungen

Wir betrachten nun *Subgradient-Descent*

$$x_{t+1} = x_t - \gamma_t g_t, \quad g_t \in \partial f(x_t).$$

Dabei ist g_t *irgendein* Element aus $\partial f(x_t)$.

Ist f konvex, dann ist $\partial f(x) \neq \varnothing \ \forall x$, so dass immer mindestens ein g_t existiert.

Ist f konvex und in x_t differenzierbar, dann gilt

$$\partial f(x_t) = \{f'(x_t)\}$$

und somit

$$g_t = f'(x_t).$$

Wir lassen auch variable Schrittweite γ_t zu.

Unter diesen Voraussetzungen wollen wir analog zu Gradient-Descent die Konvergenzeigenschaften untersuchen. Viele Ergebnisse lassen sich direkt übertragen, indem man einfach f_t' durch g_t ersetzt.

Für Schrittweite γ erhalten wir damit

$$g_t^T (x_t - x_*) = \frac{1}{2\gamma}\left(\gamma^2 \|g_t\|_2^2 + \|x_t - x_*\|_2^2 - \|x_{t+1} - x_*\|_2^2\right).$$

Da g_t Subgradient in x_t ist, gilt

$$f_* \geq f_t + g_t^T (x_* - x_t)$$

und somit

$$\begin{aligned} f_t - f_* &\leq g_t^T (x_t - x_*) \\ &= \frac{1}{2\gamma}\left(\gamma^2 \|g_t\|_2^2 + \|x_t - x_*\|_2^2 - \|x_{t+1} - x_*\|_2^2\right). \end{aligned}$$

1.4 Lipschitz-Stetigkeit

Lipschitz-Stetigkeit von f, d.h.

$$\|f(y) - f(x)\| \leq L_f \|y - x\| \quad \forall x, y \in \mathbb{R}^d$$

ist äquivalent zu

$$\|g\| \leq L_f \quad \forall g \in \partial f(x) \quad \forall x \in \mathbb{R}^d.$$

Damit erhalten wir aus der letzten Ungleichung im vorherigen Abschnitt

$$f_t - f_* \leq \frac{1}{2\gamma}\left(\gamma^2 L_f^2 + \|x_t - x_*\|_2^2 - \|x_{t+1} - x_*\|_2^2\right)$$

und somit

$$\sum_{t=0}^{T-1}(f_t - f_*) \leq \frac{\gamma}{2}TL_f^2 + \frac{1}{2\gamma}\big(\underbrace{\|x_0 - x_*\|_2^2}_{e_0^2} - \underbrace{\|x_T - x_*\|_2^2}_{\geq 0}\big)$$
$$\leq \frac{\gamma T L_f^2}{2} + \frac{e_0^2}{2\gamma},$$

was identisch ist mit der Abschätzung bei Gradient-Descent.

Somit erhalten wir in diesem Fall auch die selbe Konvergenzaussage wie bei Gradient-Descent.

Satz: $f : \mathbb{R}^d \to \mathbb{R}$, konvex, L-stetig mit Konstante L_f und es existiere $x_* = \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$. Mit $\gamma = \frac{c}{T^\omega}$, $\omega \in (0,1)$, gilt

$$\min_{t=0,\dots,T-1}(f_t - f_*) \leq \frac{1}{T}\sum_{t=0}^{T-1}(f_t - f_*)$$
$$= \mathcal{O}\Big(\big(\frac{1}{T}\big)^{\min(\omega,1-\omega)}\Big) \quad \text{für} \quad T \to \infty.$$

Die optimale Ordnung ist $\frac{1}{2}$ für $\omega = \frac{1}{2}$. Mit $e_0 = \|x_0 - x_*\|$, $\gamma = \frac{e_0}{L_f\sqrt{T}}$ gilt außerdem

$$\min_{t=0,\dots,T-1}(f_t - f_*) \leq \frac{1}{T}\sum_{t=0}^{T-1}(f_t - f_*) \leq \frac{L_f e_0}{\sqrt{T}}.$$

1.5 L-Glattheit

Bei differenzierbarem f hatten wir unter zusätzlichen Voraussetzungen (L-Glattheit, μ-Konvexität) höhere Konvergenzraten nachweisen können.

Bei L-Glattheit stoßen wir hier auf ein Problem. Wie das folgende Lemma zeigt, sind L-glatte Funktionen mit existierenden Subgradienten automatisch differenzierbar.

Lemma: $f : \mathbb{R}^d \supset \operatorname{dom}(f) \to \mathbb{R}$, $\operatorname{dom}(f)$ offen und in $x \in \operatorname{dom}(f)$ gelte $\partial f(x) \neq \emptyset$. Gibt es ein $L \geq 0$ so dass für $g_x \in \partial f(x)$

$$f(y) \leq f(x) + g_x^T(y-x) + \frac{L}{2}\|y-x\|_2^2 \quad \forall y \in \operatorname{dom}(f)$$

gilt, dann ist f differenzierbar in x und $f'(x) = g_x$.

Beweis:

- wegen $g_x \in \partial f(x)$ gilt
$$f(y) \geq f(x) + g_x^T(y-x) \quad \forall y \in \operatorname{dom}(f)$$
und es folgt
$$0 \leq f(y) - \big(f(x) + g_x^T(y-x)\big) \leq \frac{L}{2}\|y-x\|_2^2$$
- also ist für $y \to x$
$$\big|f(y) - \big(f(x) + g_x^T(y-x)\big)\big| = \mathcal{O}(\|y-x\|)$$
und somit f in x differenzierbar mit Ableitung $f'(x) = g_x$

□

Verlangen wir also L-Glattheit für subdifferenzierbare Funktionen (z.B. f konvex), so landen wir wieder beim differenzierbarem Fall.

1.6 μ-Konvexität

μ-Konvexität lässt sich einfach auf den subdifferenzierbare Funktionen verallgemeinern.

Definition: $f : \mathbb{R}^d \supset \text{dom}(f) \to \mathbb{R}$ heißt μ-konvex, falls

$$f(y) \geq f(x) + g^T(y-x) + \frac{\mu}{2}\|y-x\|_2^2 \quad \forall g \in \partial f(x) \quad \forall x, y \in \text{dom}(f).$$

Wie im letzten Kapitel erklärt, müssen wir bei nicht differenzierbarem f auf L-Glattheit verzichten, was einige Schwierigkeiten verursachen wird.

Die Probleme die sich dabei ergeben werden wir am folgenden Beispiel näher untersuchen.

Beispiel:

- für die Funktion $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = e^{|x|} = \begin{cases} e^{-x} & x < 0 \\ e^x & 0 \leq x \end{cases}$$

  erhalten wir als Subgradient

$$\partial f(x) = \begin{cases} \text{sign}(x)\ e^{|x|} & x \neq 0 \\ [-1, 1] & x = 0 \end{cases}$$

- f ist μ-konvex mit $\mu = 1$ (Übung)
- μ-Konvexität liefert nur eine Schranke nach unten, d.h. f kann sehr stark wachsen und somit kann ∂f sehr große Werte annehmen
- ist dies der Fall, dann können bei Subgradient-Descent “Overshoots” auftreten
- um dies zu kompensieren und Konvergenz zu sichern, muss die Schrittweite γ sehr klein gewählt werden
- für “brave” Funktionen f verursacht das aber unnötigen Aufwand
- als Konsequenz muss die Schrittweite entsprechend an f angepasst werden

Beim einfachen Gradient-Descent hat die L-Glattheit (als Beschränkung nach oben) dieses Problem beseitigt. Da L-Glattheit hier nicht zur Verfügung steht, werden wir unter der zusätzlichen Annahme

$$\|g_t\| \leq B \quad \forall t$$

und t-abhängiger Schrittweite γ_t das folgende Konvergenz-Resultat für Subgradient-Descent beweisen.

Satz: f sei μ-konvex und es existiere x_*. Für

$$\gamma_t = \frac{2}{\mu(t+1)}$$

erhalten wir für Subgradient-Descent

$$f\Big(\underbrace{\frac{2}{T(T+1)}\sum_{t=1}^{T} t x_t}_{\text{Konvexkombination der } x_t}\Big) - f_* \leq \frac{2B^2}{\mu(T+1)}$$

mit

$$B = \max_{t=1,\dots,T} \|g_t\|_2.$$

Beweis:

- aus den Vorüberlegungen wissen wir

$$g_t^T(x_t - x_*) = \frac{1}{2\gamma_t}\left(\gamma_t^2\|g_t\|_2^2 + \|x_t - x_*\|_2^2 - \|x_{t+1} - x_*\|_2^2\right)$$

- aus der μ-Konvexität von f folgt mit $y = x_*$, $x = x_t$

$$f_* \geq f_t + g_t^T(x_* - x_t) + \frac{\mu}{2}\|x_* - x_t\|_2^2$$

also

$$f_t - f_* \leq g_t^T(x_t - x_*) - \frac{\mu}{2}\|x_t - x_*\|_2^2$$

und somit

$$\begin{aligned} f_t - f_* &\leq \frac{1}{2\gamma_t}\big(\gamma_t^2\|g_t\|_2^2 + \|x_t - x_*\|_2^2 - \|x_{t+1} - x_*\|_2^2\big) - \frac{\mu}{2}\|x_t - x_*\|_2^2 \\ &\leq \frac{B^2}{2}\gamma_t + \frac{1}{2}\big(\frac{1}{\gamma_t} - \mu\big)\|x_t - x_*\|_2^2 - \frac{1}{2\gamma_t}\|x_{t+1} - x_*\|_2^2 \end{aligned}$$

- multiplizieren wir mit t und setzen wir wieder $e_t = \|x_t - x_*\|_2$, so erhalten wir

$$t(f_t - f_*) \leq \frac{t}{2}\Big(B^2\gamma_t + \underbrace{\big(\frac{1}{\gamma_t} - \mu\big)}_{\alpha_t}\ e_t^2 - \underbrace{\frac{1}{\gamma_t}}_{\beta_t}\ e_{t+1}^2\Big)$$

- Summation über t liefert

$$\begin{aligned} \sum_{t=1}^{T} t(f_t - f_*) &= \sum_{t=0}^{T} t(f_t - f_*) \\ &\leq \sum_{t=0}^{T} \frac{t}{2}\Big(B^2\gamma_t + \alpha_t e_t^2 - \beta_t e_{t+1}^2\Big) \\ &= \frac{B^2}{2}\sum_{t=1}^{T} t\gamma_t + \frac{1}{2}\Big(\sum_{t=0}^{T} t\alpha_t e_t^2 - \sum_{t=0}^{T} t\beta_t e_{t+1}^2\Big) \\ &= \frac{B^2}{2}\sum_{t=1}^{T} t\gamma_t + \frac{1}{2}\Big(\sum_{t=0}^{T-1} (t+1)\alpha_{t+1} e_{t+1}^2 - \sum_{t=0}^{T} t\beta_t e_{t+1}^2\Big) \\ &= \frac{B^2}{2}\sum_{t=1}^{T} t\gamma_t + \frac{1}{2}\Big(\sum_{t=0}^{T-1} \big((t+1)\alpha_{t+1} - t\beta_t\big)e_{t+1}^2 - \beta_T e_{T+1}^2\Big) \\ &\leq \frac{B^2}{2}\sum_{t=1}^{T} t\gamma_t + \frac{1}{2}\Big(\sum_{t=0}^{T-1} \big((t+1)\alpha_{t+1} - t\beta_t\big)e_{t+1}^2\Big) \end{aligned}$$

- α_t und β_t hängen nur von γ_t und μ ab
- kann man γ_t so wählen, dass

$$(t+1)\alpha_{t+1} - t\beta_t \leq 0 \quad \forall t \geq 0$$

gilt, dann kann der zweite Summand auf der rechten Seite mit 0 nach oben abgeschätzt werden

- für

$$\gamma_t = \frac{2}{\mu(t+1)}$$

ist

$$\begin{aligned} (t+1)\alpha_{t+1} - t\beta_t &= (t+1)\big(\frac{1}{\gamma_{t+1}} - \mu\big) - t\frac{1}{\gamma_t} \\ &= (t+1)\big(\frac{\mu(t+2)}{2} - \mu\big) - t\frac{\mu(t+1)}{2} \\ &= \frac{\mu}{2}(t+1)t - t\frac{\mu}{2}(t+1) \\ &= 0 \end{aligned}$$

und damit

$$\sum_{t=1}^{T} t(f_t - f_*) \le \frac{B^2}{2} \sum_{t=1}^{T} t\gamma_t$$

- wegen

  $$\frac{2}{T(T+1)} \sum_{t=1}^{T} t = 1,$$

  der Konvexität von f und der Jensen-Ungleichung erhalten wir

  $$\begin{aligned} f\Big(\frac{2}{T(T+1)} \sum_{t=1}^{T} t x_t\Big) - f_* &\le \Big(\frac{2}{T(T+1)} \sum_{t=1}^{T} t f_t\Big) - f_* \\ &= \frac{2}{T(T+1)} \sum_{t=1}^{T} t(f_t - f_*) \\ &\le \frac{B^2}{T(T+1)} \sum_{t=1}^{T} t\gamma_t \end{aligned}$$

- mit

  $$\gamma_t = \frac{2}{\mu(t+1)}$$

  folgt

  $$\sum_{t=1}^{T} t\gamma_t = \frac{2}{\mu} \sum_{t=1}^{T} \frac{t}{t+1} \le \frac{2}{\mu} T$$

  und somit schließlich

  $$f\Big(\frac{2}{T(T+1)} \sum_{t=1}^{T} t x_t\Big) - f_* \le \frac{2B^2}{\mu(T+1)}$$

□

Bemerkung:

- in der oberen Schranke steckt x_0 nicht explizit drin, geht aber implizit über B ein ($\|g_t\|_2 \le B \ \forall t$)
- für f differenzierbar, μ-konvex und L-glatt hatten wir bei Gradient-Descent die Komplexität $\mathcal{O}\big(\log(\frac{1}{\varepsilon})\big)$
- für f μ-konvex, $\|g_t\|_2 \le B \ \forall t$ bekommen wir bei Subgradient-Descent nur die Komplexität $\mathcal{O}\big(\frac{1}{\varepsilon}\big)$, d.h. aus fehlender Glattheit von f kann wieder langsamere Konvergenz folgen

1.7 Zusammenfassung

Ist f konvex und nicht differenzierbar, dann benutzt man bei Gradient-Descent statt des (eventuell nicht existierenden) Gradienten einen Subgradienten.

Für das *nicht restringierte Optimierungsproblem* haben wir folgendes Konvergenzverhalten nachgewiesen:

- f konvex und Lipschitz-stetig, $\gamma = \frac{c}{\sqrt{T}}$, $c > 0$:

  $$\min_{t=0,\dots,T-1} (f_t - f_*) \le \frac{L_f e_0}{\sqrt{T}}$$

  also

  $$\min_{t=0,\dots,T-1} (f_t - f_*) \le \varepsilon \quad \Rightarrow \quad T = \mathcal{O}\big(\frac{1}{\varepsilon^2}\big)$$

- f μ-konvex mit $\mu > 0$ und $\|g_t\| \leq B$, $\gamma_t = \frac{2}{\mu(t+1)}$

  $$f\Big(\frac{2}{T(T+1)}\sum_{t=1}^{T} t x_t\Big) - f_* \leq \frac{2B^2}{\mu(T+1)},$$

  also

  $$f\Big(\frac{2}{T(T+1)}\sum_{t=1}^{T} t x_t\Big) - f_* \leq \varepsilon \quad \Rightarrow \quad T = \mathcal{O}\Big(\frac{1}{\varepsilon}\Big)$$

Im ersten Fall benötigen wir für Genauigkeit ε den selben asymptotischen Aufwand wie bei differenzierbarem f.

Im zweiten Fall steigt der Aufwand von $\mathcal{O}\Big(\log\big(\frac{1}{\varepsilon}\big)\Big)$ auf $\mathcal{O}\Big(\frac{1}{\varepsilon}\Big)$, d.h. durch die reduzierten Glattheitsanforderungen an f reduziert sich hier die Konvergenzgeschwindigkeit.