14_Stochastic_Gradient_Descent

Martin Reißel

27. Juni 2022

Inhaltsverzeichnis

1 Stochastic Gradient Descent

1.1 Überblick

In viele Anwendungen hat f die Struktur

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \text{ differenzierbar},$$

z.B. wenn f_i die Loss-Funktion zum i-ten Trainings-Datensatz ist.

Bei einfachem Gradient-Descent muss in jedem Schritt

$$f'(x) = \frac{1}{n}\sum_{i=1}^{n} f_i'(x),$$

berechnet werden. Dies wird nun wie folgt vereinfacht:

- wähle in jedem Schritt zufällig (gleichverteilt) ein $i_t \in \{1, \ldots, n\}$ aus
- setze
$$x_{t+1} = x_t - \gamma_t f_{i_t}'(x_t)$$

Man bezeichnet

$$g_t = g(i_t, x_t) = f_{i_t}'(x_t)$$

als *stochastischen Gradienten*. Der Aufwand pro Schritt wird im Vergleich zum einfachen Gradient-Descent-Verfahren um den Faktor n reduziert.

1.2 Vorüberlegungen

Eine direkte Übertragung der Konvergenzanalysen des Gradient-Descent-Verfahrens funktioniert zunächst nicht.

Bei Gradient-Descent hatten wir aus der Konvexitätsbedingung

$$f(y) \geq f(x) + f'(x)(y - x)$$

die Ungleichung

$$f_t - f_* \leq f_t'(x_t - x_*)$$

abgeleitet und die weiteren Untersuchungen darauf aufgebaut.

Bei Stochastic-Gradient-Descent geht das nicht, da $g_t \neq f'(x_t)$. Wir können aber zeigen, dass die Ungleichung für Erwartungswerte gilt:

- ist $i_t \in \{1, \ldots, n\}$ eine gleichverteilte Zufallsvariable, dann gilt

  $$\mathbb{P}(i_t = k) = \frac{1}{n} \quad \forall k \in \{1, \ldots, n\}$$

  und somit

  $$\begin{aligned} \mathbb{E}_{i_t}(g_t) &= \sum_{k=1}^{n} f_{i_t}'(x_t)\mathbb{P}(i_t = k) \\ &= \frac{1}{n}\sum_{k=1}^{n} f_{i_t}'(x_t) \\ &= f'(x_t) \end{aligned}$$

- g_t ist also ein erwartungstreuer Schätzer von $f'(x_t)$

Wir beginnen mit einer grundlegenden Abschätzung für Stochastic-Gradient-Descent

$$x_{t+1} = x_t - \gamma \underbrace{g(i_t, x_t)}_{g_t}.$$

Lemma: Ist f' Lipschitz-stetig mit Konstante L, dann gilt

$$\mathbb{E}_{i_t}(f_{t+1}) - f_t \leq -\gamma f_t' \, \mathbb{E}_{i_t}(g_t) + \frac{1}{2}\gamma^2 L \, \mathbb{E}_{i_t}(\|g_t\|_2^2).$$

Beweis:

- ist f' Lipschitz-stetig dann gilt (siehe Gradient-Descent)

  $$f(y) \leq f(x) + f'(x)(y - x) + \frac{1}{2}L\|y - x\|_2^2$$

- mit $y = x_{t+1}$, $x = x_t$ erhalten wir

  $$f_{t+1} - f_t \leq f_t'(x_{t+1} - x_t) + \frac{1}{2}L\|x_{t+1} - x_t\|_2^2$$

- wegen $x_{t+1} - x_t = -\gamma g_t$ ist

  $$f_{t+1} - f_t \leq -\gamma f_t' g_t + \frac{1}{2}\gamma^2 L\|g_t\|_2^2$$

- bilden wir auf beiden Seiten $\mathbb{E}_{i_t}$ und beachten wir, dass f_t, f_t' nicht von i_t abhängen, dann folgt

  $$\mathbb{E}_{i_t}(f_{t+1}) - f_t \leq -\gamma f_t' \, \mathbb{E}_{i_t}(g_t) + \frac{1}{2}\gamma^2 L \, \mathbb{E}_{i_t}(\|g_t\|_2^2)$$

□

Bemerkung:

- der erwartete Abstieg ist beschränkt durch den Erwartungswert der Richtungsableitung von f in Richtung g_t

$$f_t' \, \mathbb{E}_{i_t}(g_t) = \mathbb{E}_{i_t}(g_t)^T f_t'$$

  sowie

$$\mathbb{E}_{i_t}(\|g_t\|_2^2)$$

- wie man zu x_t gekommen ist, spielt keine Rolle (Markov-Eigenschaft)
- ist g_t ein erwartungstreuer Schätzer, d.h. $\mathbb{E}_{i_t}(g_t) = f_t'$, dann folgt

$$\begin{aligned}\mathbb{E}_{i_t}(f_{t+1}) - f_t &\leq -\gamma f_t' \, \mathbb{E}_{i_t}(g_t) + \frac{1}{2}\gamma^2 L \, \mathbb{E}_{i_t}(\|g_t\|_2^2) \\ &\leq -\gamma \|f_t'\|_2^2 + \frac{1}{2}\gamma^2 L \, \mathbb{E}_{i_t}(\|g_t\|_2^2)\end{aligned}$$

- kann $\mathbb{E}_{i_t}(\|g_t\|_2^2)$ "deterministisch" beschränkt werden, dann erhalten wir wieder einen hinreichenden Abstieg

Definition: Es sei

$$\mathbb{V}_{i_t} = \mathbb{E}_{i_t}(\|g_t\|_2^2) - \|\mathbb{E}_{i_t}(g_t)\|_2^2.$$

Für den Rest des Kapitels treffen wir folgende Annahmen:

- $x_t \in X$, X offen, $f|_X \geq \bar{f}$
- $\exists c_G \geq c > 0$ so dass $\forall t$ gilt

$$f_t \, \mathbb{E}_{i_t}(g_t) \geq c\|f_t'\|_2^2, \quad \|\mathbb{E}_{i_t}(g_t)\|_2 \leq c_G \|f_t'\|_2$$

- $\exists M, M_V \geq 0$ mit

$$\mathbb{V}_{i_t} \leq M + M_V \|f_t'\|_2^2$$

Bemerkung: Ist g_t ein erwartungstreuer Schätzer von f_t', dann ist $c = c_G = 1$.

Aus der Annahme folgt nun

$$\begin{aligned}\mathbb{E}_{i_t}(\|g_t\|_2^2) &= \mathbb{V}_{i_t} + \|\mathbb{E}_{i_t}(g_t)\|_2^2 \\ &\leq M + M_V \|f_t'\|_2^2 + c_G^2 \|f_t'\|_2^2 \\ &\leq M + \underbrace{(M_V + c_G^2)}_{M_G \geq 0} \|f_t'\|_2^2.\end{aligned}$$

Kombiniert mit dem letzten Lemma erhalten wir

$$\begin{aligned}\mathbb{E}_{i_t}(f_{t+1}) - f_t &\leq -\gamma f_t' \, \mathbb{E}_{i_t}(g_t) + \frac{1}{2}\gamma^2 L \, \mathbb{E}_{i_t}(\|g_t\|_2^2) \\ &\leq -\gamma c \|f_t'\|_2^2 + \frac{1}{2}\gamma^2 L\big(M + M_G \|f_t'\|_2^2\big) \\ &= -\gamma\big(c - \frac{\gamma L}{2} M_G\big)\|f_t'\|_2^2 + \frac{1}{2}\gamma^2 LM\end{aligned}$$

und somit

Lemma:

$$\mathbb{E}_{i_t}(f_{t+1}) - f_t \leq -\gamma\big(c - \frac{\gamma L}{2} M_G\big)\|f_t'\|_2^2 + \frac{1}{2}\gamma^2 LM$$

Bemerkung:

- die rechte Seite der Abschätzung ist deterministisch
- die Schranke für $\mathbb{E}_{i_t}(f_{t+1}) - f_t$ hängt nur von x_t und i_t ab und *nicht* von früheren x_s, $s < t$
- ist γ klein genug, dann ist $\mathbb{E}_{i_t}(f_{t+1}) - f_t < 0$

1.3 Konvergenz

f sei jetzt differenzierbar, L-glatt und μ-konvex und zusätzlich soll x_* mit $f_* = f(x_*) = \inf_x f(x)$ existieren. Wir werden nun zeigen, dass wir unter diesen Voraussetzungen bei konstanter Schrittweite γ zwar keine Konvergenz erhalten, aber zumindest in eine Umgebung von x_* gelangen, deren Größe asymptotisch proportional zu γ ist.

Wir verwenden im Folgenden die Notation

$$\mathbb{E}(f_t) = \mathbb{E}_{i_0} \dots \mathbb{E}_{i_{t-1}}(f_t).$$

Satz: Es sei f differenzierbar, L-glatt, μ-konvex und es existiere x_* mit $f_* = f(x_*) = \inf_x f(x)$. Ist

$$0 < \gamma \leq \frac{c}{LM_G}$$

dann gilt

$$\begin{aligned}\mathbb{E}(f_t - f_*) &\leq \frac{\gamma LM}{2\mu c} + (1 - \gamma\mu c)^{t-1}\big(\mathbb{E}(f_1 - f_*) - \frac{\gamma LM}{2\mu c}\big) \\ &\xrightarrow{t\to\infty} \frac{\gamma LM}{2\mu c}.\end{aligned}$$

Beweis:

- aus dem letzten Lemma im vorherigen Abschnitt wissen wir

$$\begin{aligned}\mathbb{E}_{i_t}(f_{t+1} - f_t) &= \mathbb{E}_{i_t}(f_{t+1}) - f_t \\ &\leq -\gamma\big(c - \frac{\gamma L}{2} M_G\big)\|f_t'\|_2^2 + \frac{1}{2}\gamma^2 LM\end{aligned}$$

- wegen

$$0 < \gamma \leq \frac{c}{LM_G}$$

ist

$$\frac{\gamma L}{2} M_G \leq \frac{c}{2},$$

und deshalb

$$\mathbb{E}_{i_t}(f_{t+1}) - f_t \leq -\frac{\gamma c}{2}\|f_t'\|_2^2 + \frac{1}{2}\gamma^2 LM$$

- oben hatten wir gesehen, dass für μ-konvexe Funktionen auf $\mathbb{R}^d$

$$f(x) - f_* \leq \frac{1}{2\mu}\|f'(x)\|_2^2$$

gilt, also

$$\|f_t'\|_2^2 \geq 2\mu(f_t - f_*)$$

und somit

$$\mathbb{E}_{i_t}(f_{t+1} - f_*) - (f_t - f_*) \leq -\gamma\mu c(f_t - f_*) + \frac{1}{2}\gamma^2 LM$$

- nun bilden wir auf beiden Seiten die Erwartungswerte $\mathbb{E}_{i_1} \dots \mathbb{E}_{i_{t-1}}$ und erhalten

$$\mathbb{E}(f_{t+1} - f_*) - \mathbb{E}(f_t - f_*) \leq -\gamma\mu c\mathbb{E}(f_t - f_*) + \frac{1}{2}\gamma^2 LM$$

also

$$\mathbb{E}(f_{t+1} - f_*) \leq (1 - \gamma\mu c)\mathbb{E}(f_t - f_*) + \frac{1}{2}\gamma^2 LM$$

- subtrahiert man $\frac{\gamma LM}{2\mu c}$ auf beiden Seiten, so ergibt sich

$$\begin{aligned}\mathbb{E}(f_{t+1} - f_*) - \frac{\gamma LM}{2\mu c} &\leq (1 - \gamma\mu c)\mathbb{E}(f_t - f_*) + \frac{1}{2}\gamma^2 LM - \frac{\gamma LM}{2\mu c} \\ &= (1 - \gamma\mu c)\big(\mathbb{E}(f_t - f_*) - \frac{\gamma LM}{2\mu c}\big)\end{aligned}$$

- wegen $\mu \leq L$ und $M_G \geq c_G^2 \geq c^2$ ist

$$0 < \gamma\mu c \leq \frac{\mu c^2}{LM_G} \leq \frac{c^2}{M_G} \leq 1$$

□

Ohne "Rauschen" ist $M = 0$ und

$$\mathbb{E}(f_t - f_*) \leq (1 - \gamma\mu c)^{t-1}\mathbb{E}(f_1 - f_*)$$

so dass wir eine Komplexität von $\mathcal{O}\big(\log(\frac{1}{\varepsilon})\big)$ erhalten.

Mt "Rauschen" haben wir

$$\mathbb{E}(f_t - f_*) \leq \underbrace{\frac{\gamma LM}{2\mu c}}_{\text{fix}} + \underbrace{(1 - \gamma\mu c)^{t-1}\big(\mathbb{E}(f_1 - f_*) - \frac{\gamma LM}{2\mu c}\big)}_{\text{geometrische Reduktion}},$$

d.h. wir erreichen

$$\mathbb{E}(f_t - f_*) \leq \frac{\gamma LM}{2\mu c} + \varepsilon$$

in

$$t = \mathcal{O}\big(\log(\frac{1}{\varepsilon})\big)$$

Schritten.

$\frac{\gamma LM}{2\mu c}$ wird klein, wenn γ klein wird, wobei dann aber $1 - \gamma\mu c$ nahe bei 1 liegt, so dass die geometrische Reduktion nur sehr langsam ist.

Dies führt auf die Idee des Restarts:

- starte mit einer Schrittweite $\gamma^{(1)}$ und iteriere bis

$$\mathbb{E}(f_t - f_*) \leq \frac{\gamma^{(1)} LM}{2\mu c} + \varepsilon^{(1)}$$

- verkürze die Schrittweite auf $\gamma^{(2)} < \gamma^{(1)}$ und iteriere bis die rechte Seite hinreichend klein ist

Das kann man soweit ausbauen, dass man in jedem Schritt die Schrittweite γ anpasst.

Satz: Es sei f differenzierbar, L-glatt, μ-konvex. Es existiere ein x_* mit $f_* = f(x_*) = \inf_x f(x)$ und es sei

$$\gamma_t = \frac{\beta}{\gamma + t}, \quad \beta > \frac{1}{\mu c}, \quad \gamma > 0, \quad \gamma_1 < \frac{c}{LM_G}.$$

Dann gilt

$$\mathbb{E}(f_t - f_*) \leq \frac{\nu}{\gamma + t}$$

mit

$$\nu = \max\Big(\frac{\beta^2 LM}{2(\beta\mu c - 1)}, (\gamma + 1)\ \mathbb{E}(f_1 - f_*)\Big).$$

Beweis:

- aus den Voraussetzungen folgt
$$\gamma_t L M_G \leq \gamma_1 L M_G \leq c \quad \forall t$$
- mit $\gamma = \gamma_t$ folgt aus dem letzten Lemma
$$\begin{aligned}\mathbb{E}_{i_t}(f_{t+1}) - f_t &\leq -\gamma_t\big(c - \frac{\gamma_t L}{2} M_G\big)\|f_t'\|_2^2 + \frac{1}{2}\gamma_t^2 LM \\ &= \big(-\gamma_t c + \gamma_t \frac{\gamma_t L M_G}{2}\big)\|f_t'\|_2^2 + \frac{1}{2}\gamma_t^2 LM \\ &\leq -\frac{1}{2}\gamma_t c\|f_t'\|_2^2 + \frac{1}{2}\gamma_t^2 LM\end{aligned}$$
- wegen der μ-Konvexität von f auf $\mathbb{R}^d$ gilt wieder
$$\|f_t'\|_2^2 \geq 2\mu(f_t - f_*)$$
und somit
$$\mathbb{E}_{i_t}(f_{t+1}) - f_t \leq -\gamma_t \mu c(f_t - f_*) + \frac{1}{2}\gamma_t^2 LM,$$
bzw.
$$\mathbb{E}_{i_t}(f_{t+1}) - f_* - (f_t - f_*) \leq -\gamma_t \mu c(f_t - f_*) + \frac{1}{2}\gamma_t^2 LM$$
- nun bilden wir auf beiden Seiten die Erwartungswerte $\mathbb{E}_{i_1} \dots \mathbb{E}_{i_{t-1}}$ und erhalten
$$\mathbb{E}(f_{t+1} - f_*) \leq (1 - \gamma_t \mu c)\,\mathbb{E}(f_t - f_*) + \frac{1}{2}\gamma_t^2 LM$$
- per Induktion können wir nun die Aussage des Satzes beweisen
- für $t = 1$ folgt die Behauptung direkt aus der Definition von ν
- für den Schritt $t \to t+1$ erhalten wir
$$\begin{aligned}\mathbb{E}(f_{t+1} - f_*) &\leq (1 - \gamma_t \mu c)\,\mathbb{E}(f_t - f_*) + \frac{1}{2}\gamma_t^2 LM \\ &\leq \big(1 - \frac{\beta}{\gamma + t}\mu c\big)\,\mathbb{E}(f_t - f_*) + \frac{1}{2}\frac{\beta^2}{(\gamma + t)^2} LM \\ &\leq \big(1 - \frac{\beta}{\gamma + t}\mu c\big)\,\frac{\nu}{\gamma + t} + \frac{1}{2}\frac{\beta^2}{(\gamma + t)^2} LM \\ &= \frac{(\gamma + t - \beta\mu c)\nu}{(\gamma + t)^2} + \frac{\beta^2 LM}{2(\gamma + t)^2} \\ &= \frac{\gamma + t - 1}{(\gamma + t)^2}\nu - \frac{\beta\mu c - 1}{(\gamma + t)^2}\nu + \frac{\beta^2 LM}{2(\gamma + t)^2} \\ &= \frac{\gamma + t - 1}{(\gamma + t)^2}\nu + \frac{\beta^2 LM - 2(\beta\mu c - 1)\nu}{2(\gamma + t)^2}\end{aligned}$$
- nach Definition von ν gilt
$$\nu \geq \frac{\beta^2 LM}{2(\beta\mu c - 1)}$$
also
$$\beta^2 LM - 2(\beta\mu c - 1)\nu \leq 0$$

und somit

$$\begin{aligned}\mathbb{E}(f_{t+1} - f_*) &\leq \frac{\gamma + t - 1}{(\gamma + t)^2}\nu \\ &\leq \frac{(\gamma + t - 1)(\gamma + t + 1)}{(\gamma + t)^2}\frac{\nu}{\gamma + t + 1} \\ &= \frac{(\gamma + t)^2 - 1}{(\gamma + t)^2}\frac{\nu}{\gamma + t + 1} \\ &\leq \frac{\nu}{\gamma + t + 1}\end{aligned}$$

□

1.4 Zusammenfassung

Ist $f : \mathbb{R}^d \to \mathbb{R}$ differenzierbar, L-glatt und μ-konvex mit $f_* = f(x_*) = \inf_x f(x)$ dann haben wir für Stochastic-Gradient-Descent

$$x_{t+1} = x_t - \gamma_t f'_{i_t}(x_t)$$

folgende Ergebnisse gezeigt:

- ist γ konstant, dann ist
$$\mathbb{E}(f_t) - f_* \leq c(\gamma) + \varepsilon$$
in $\mathcal{O}(\log \frac{1}{\varepsilon})$ Schritten
- ist $\gamma \sim \frac{1}{t}$, dann ist
$$\mathbb{E}(f_t) - f_* \leq \varepsilon$$
mit $t = \mathcal{O}(\frac{1}{\varepsilon})$
- Gradient-Descent liefert für f L-glatt und μ-konvex , $0 < \gamma \leq \frac{1}{L}$
$$f_t - f_* \leq \varepsilon$$
in $t = \mathcal{O}\Big(\log\big(\frac{1}{\varepsilon}\big)\Big)$ Schritten

Bei Stochastic-Gradient-Descent benötigen wir also deutlich mehr Schritte, wobei jeder einzelne Schritt wegen der vereinfachten Gradientenberechnung f'_{i_t} sehr viel weniger Aufwand verursacht.

Proximal- bzw. Subgradienten-Verfahren können analog “umgebaut” werden.