13_Proximal_Gradient_Descent

Martin Reißel

27. Juni 2022

Inhaltsverzeichnis

1 Proximal Gradient Descent

1.1 Überblick

Bei Subgradient-Descent haben wir gesehen, dass fehlende Differenzierbarkeit von f kein Problem ist, solange Subgradienten existieren. Allerdings reduziert sich bei fehlender Glattheit die Konvergenzgeschwindigkeit.

Hat f zusätzlich eine spezielle Struktur, so kann man Gradient-Descent so modifizieren, dass man Konvergenzraten wie im C^1-Fall erhält.

Wir untersuchen hier nicht restringierte Probleme, restringierte lassen sich bei Proximal-Gradient-Descent durch geeignete Wahl der Zielfunktion auf restringierte zurück führen.

1.2 Grundlagen

Für $\gamma > 0$ und festes z ist die Funktion

$$q(y) = \frac{1}{2\gamma}\|y - z\|_2^2$$

strikt konvex mit globalem Minimum $y_* = z$.

Für f differenzierbar kann man deshalb einen Gradient-Descent-Schritt

$$x_{t+1} = x_t - \gamma f_t'$$

auch schreiben kann als

$$x_{t+1} = \underset{y}{\operatorname{argmin}}\Big(\frac{1}{2\gamma}\big\|y - (x_t - \gamma f_t')\big\|_2^2\Big).$$

Wir nehmen nun an, dass $f = g + h$ ist mit

- g differenzierbar, L-glatt und μ-konvex
- h “nur” konvex

Beispiel: Bei Lasso ist

$$f(w) = \underbrace{\frac{1}{2m}\|Xw - y\|_2^2}_{g(w)} + \underbrace{\alpha\|w\|_1}_{h(w)}.$$

Wir benutzen jetzt die Darstellung von Gradient-Descent von oben, behandeln g "wie üblich" und hängen h "unbehandelt" an

$$\begin{aligned} x_{t+1} &= \underset{y}{\operatorname{argmin}}\Big(\frac{1}{2\gamma}\|y - (x_t - \gamma g_t')\|_2^2 + h(y)\Big) \\ &= \underset{y}{\operatorname{argmin}}\Big(\frac{1}{2}\|y - (x_t - \gamma g_t')\|_2^2 + \gamma h(y)\Big). \end{aligned}$$

Das führt uns zu folgender Notation.

Definition:

- für h konvex, $\gamma > 0$ ist der *Proximal-Operator* definiert als

$$\operatorname{prox}(x) = \underset{y}{\operatorname{argmin}}\Big(\frac{1}{2}\|y - x\|_2^2 + \gamma h(y)\Big)$$

- für $f = g + h$, g differenzierbar, h konvex, ist *Proximal-Gradient-Descent* gegeben durch

$$x_{t+1} = \operatorname{prox}(y_{t+1}), \quad y_{t+1} = x_t - \gamma g_t'$$

Bemerkung:

- man kann zeigen, dass für h konvex, halbstetig von unten, "proper" ($h(x) > -\infty \ \forall x$, h nicht identisch $+\infty$) der Proximal-Operator wohldefiniert ist (Eindeutigkeit ist klar, da $\frac{1}{2}\|y - x\|_2^2 + \gamma h(y)$ strikt konvex ist, die sonstigen Voraussetzungen an h braucht man zum Nachweis der Existenz)
- Proximal-Gradient-Descent ähnelt Projected-Gradient-Descent, wobei Π durch prox ersetzt wird (Übung: Projected-Gradient-Descent ist Spezialfall von Proximal-Gradient-Descent)

1.3 Konvergenz

Die Konvergenzbeweise verlaufen ähnlich wie bei Projected-Gradient-Descent. Dazu benötigen wir Informationen über die Abbildungseigenschaften von $\operatorname{prox}(\cdot)$.

Lemma:

- ist $x_* = \underset{x}{\operatorname{argmin}} f(x)$, dann gilt $x_* = \operatorname{prox}(x_* - \gamma g_*')$ (Übung)
- $\|\operatorname{prox}(y) - \operatorname{prox}(x)\|_2 \le \|y - x\|_2 \quad \forall x, y$

Es sei nun g differenzierbar, L-glatt und μ-konvex. Wegen der L-Glattheit ist g' Lipschitz-stetig, d.h.

$$\|g'(y) - g'(x)\|_2 \le L\|y - x\|_2.$$

μ-Konvexität bedeutet

$$g(y) \ge g(x) + g'(x)(y - x) + \frac{\mu}{2}\|y - x\|_2^2 \quad \forall x, y,$$

also auch

$$g(x) \ge g(y) + g'(y)(x - y) + \frac{\mu}{2}\|x - y\|_2^2 \quad \forall x, y.$$

Addition der beiden Ungleichungen liefert

$$g(y) + g(x) \ge g(x) + g(y) + (g'(x) - g'(y))(y - x) + \mu\|y - x\|_2^2,$$

so dass

$$\big(g'(y)-g'(x)\big)(y-x) \geq \mu\|y-x\|_2^2.$$

Mit Cauchy-Schwartz folgt

$$\mu\|y-x\|_2^2 \leq \big(g'(y)-g'(x)\big)(y-x) \leq \|g'(y)-g'(x)\|_2\|y-x\|_2$$

und schließlich

$$\|g'(y)-g'(x)\|_2 \geq \mu\|y-x\|_2.$$

Für $\beta \in \mathbb{R}$ beliebig erhalten wir

$$\beta\|g'(y)-g'(x)\|_2^2 \leq \beta L^2\|y-x\|_2^2 \quad \text{für} \quad \beta \geq 0$$

bzw.

$$\beta\|g'(y)-g'(x)\|_2^2 \leq \beta\mu^2\|y-x\|_2^2 \quad \text{für} \quad \beta < 0$$

und somit

$$\beta\|g'(y)-g'(x)\|_2^2 \leq \max(\beta L^2, \beta\mu^2)\|y-x\|_2^2 \quad \forall \beta \in \mathbb{R}.$$

Außerdem ist

$$\begin{aligned}
\big(g'(y)-g'(x)\big)(y-x) &\geq \mu\|y-x\|_2^2 \\
&= \frac{L}{L+\mu}\mu\|y-x\|_2^2 + \frac{\mu}{L+\mu}\mu\|y-x\|_2^2 \\
&\geq \frac{L}{L+\mu}\mu\|y-x\|_2^2 + \frac{\mu^2}{L+\mu}\frac{1}{L^2}\|g'(y)-g'(x)\|_2^2.
\end{aligned}$$

Wegen $0 < \mu \leq L$ folgt

$$\big(g'(y)-g'(x)\big)(y-x) \geq \frac{1}{L+\mu}\|g'(y)-g'(x)\|_2^2 + \frac{L\mu}{L+\mu}\|y-x\|_2^2.$$

Damit erhalten wir das folgende Resultat.

Satz: Ist $f = g+h$, g differenzierbar, L-glatt und μ-konvex, h konvex, $x_* = \operatorname{argmin}_x f(x)$ und

$$x_{t+1} = \operatorname{prox}(x_t - \gamma g'_t),$$

dann gilt

$$\|x_t - x_*\|_2 \leq Q(\gamma)^t\|x_0 - x_*\|_2$$

mit

$$Q(\gamma) = \max(|1-\gamma L|, |1-\gamma\mu|).$$

Beweis:

- es ist

$$\begin{aligned}
\|x_{t+1}-x_*\|_2^2 &= \|\operatorname{prox}(x_t-\gamma g'_t) - \operatorname{prox}(x_*-\gamma g'_*)\|_2^2 \\
&\leq \|x_t-\gamma g'_t - (x_*-\gamma g'_*)\|_2^2 \\
&= \|x_t - x_* - \gamma(g'_t - g'_*)\|_2^2 \\
&= \|x_t-x_*\|_2^2 + \gamma^2\|g'_t-g'_*\|_2^2 - 2\gamma(g'_t-g'_*)(x_t-x_*) \\
&\leq \|x_t-x_*\|_2^2 + \gamma^2\|g'_t-g'_*\|_2^2 \\
&\quad -2\gamma\Big(\frac{1}{L+\mu}\|g'_t-g'_*\|_2^2 + \frac{L\mu}{L+\mu}\|x_t-x_*\|_2^2\Big) \\
&= \underbrace{\Big(1-\frac{2\gamma L\mu}{L+\mu}\Big)}_{\alpha}\|x_t-x_*\|_2^2 + \underbrace{\gamma\Big(\gamma-\frac{2}{L+\mu}\Big)}_{\beta}\|g'_t-g'_*\|_2^2
\end{aligned}$$

und somit

$$\begin{aligned}\|x_{t+1} - x_*\|_2^2 &\leq \alpha \|x_t - x_*\|_2^2 + \max\left(\beta L^2, \beta\mu^2\right)\|x_t - x_*\|_2^2 \\ &= \max\left(\alpha + \beta L^2, \alpha + \beta\mu^2\right)\|x_t - x_*\|_2^2\end{aligned}$$

- mit

$$\begin{aligned}\alpha + \beta L^2 &= 1 - \frac{2\gamma L\mu}{L+\mu} + \gamma\big(\gamma - \frac{2}{L+\mu}\big)L^2 \\ &= 1 + \gamma^2 L^2 - \frac{1}{L+\mu}\big(2\gamma L\mu + 2\gamma L^2\big) \\ &= 1 + \gamma^2 L^2 - \frac{2}{L+\mu}\gamma L(L+\mu) \\ &= 1 + \gamma^2 L^2 - 2\gamma L \\ &= (1-\gamma L)^2\end{aligned}$$

  und

$$\begin{aligned}\alpha + \beta\mu^2 &= 1 - \frac{2\gamma L\mu}{L+\mu} + \gamma\big(\gamma - \frac{2}{L+\mu}\big)\mu^2 \\ &= 1 + \gamma^2\mu^2 - \frac{2\gamma\mu}{L+\mu}(L+\mu) \\ &= (1-\gamma\mu)^2\end{aligned}$$

  erhalten wir schließlich

$$\|x_{t+1} - x_*\|_2^2 \leq \max\left((1-\gamma L)^2, (1-\gamma\mu)^2\right)\ \|x_t - x_*\|_2^2$$

□

1.4 Zusammenfassung

Ist $f = g + h$, g differenzierbar, L-glatt und μ-konvex, h konvex, $x_* = \mathrm{argmin}_x f(x)$, dann hat Proximal-Gradient-Descent die selbe asymptotische Konvergenzgeschwindigkeit wie im Fall dass f differenzierbar, L-glatt und μ-konvex ist. Die fehlende Glattheit von h spielt, anders als bei Subgradient-Descent, keine Rolle.

Das Berechnen von

$$\mathrm{prox}(x) = \underset{y}{\mathrm{argmin}}\big(\frac{1}{2}\|y - x\|_2^2 + \gamma h(y)\big)$$

ist in der Regel nicht trivial. Für einige spezielle h, die in der Praxis häufig auftreten, lässt sich $\mathrm{prox}(x)$ einfach bestimmen (siehe Übung).

Projected-Gradient-Descent kann als Spezialfall von Proximal-Gradient-Descent interpretiert werden (siehe Übung).