10_Gradient_Descent

Martin Reißel

27. Juni 2022

Inhaltsverzeichnis

1 Gradient Descent

1.1 Überblick

Wir betrachten das Gradient-Descent Verfahren

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

untersuchen Konvergenz, d.h.

$$f_t - f_* \xrightarrow{t\to\infty} 0$$

bzw.

$$\|x_t - x_*\| \xrightarrow{t\to\infty} 0$$

und versuchen das asymptotische Verhalten genauer zu analysieren.

Für den Rest des Kapitels setzen wir $f \in C^1(\mathbb{R}^d)$ konvex und $\gamma_t = \gamma$ konstant voraus.

1.2 Vorüberlegungen

Ist $f \in C^1(\mathbb{R}^d)$, $x_* = \operatorname{argmin}_{x\in\mathbb{R}} f(x)$, $x_{t+1} = x_t - \gamma f_t'$, dann gilt

$$\begin{aligned}\|x_{t+1} - x_*\|_2^2 &= \|x_t - x_* - \gamma f_t'\|_2^2 \\ &= \|x_t - x_*\|_2^2 + \gamma^2\|f_t'\|_2^2 - 2\gamma f_t'(x_t - x_*)\end{aligned}$$

und somit

$$f_t'(x_t - x_*) = \frac{1}{2\gamma}\left(\gamma^2\|f_t'\|_2^2 + \|x_t - x_*\|_2^2 - \|x_{t+1} - x_*\|_2^2\right).$$

Für konvexes f ist

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

und mit $y = x_*$, $x = x_t$

$$f_* \geq f_t + f_t'(x_* - x_t)$$

bzw.

$$\begin{aligned} 0 \leq f_t - f_* &\leq f_t'(x_t - x_*) \\ &= \frac{1}{2\gamma}\left(\gamma^2 \|f_t'\|_2^2 + \|x_t - x_*\|_2^2 - \|x_{t+1} - x_*\|_2^2\right). \end{aligned}$$

Aufsummiert erhält man

$$\begin{aligned} \sum_{t=0}^{T-1}(f_t - f_*) &\leq \sum_{t=0}^{T-1} f_t'(x_t - x_*) \\ &= \frac{1}{2\gamma}\sum_{t=0}^{T-1}\left(\gamma^2 \|f_t'\|_2^2 + \|x_t - x_*\|_2^2 - \|x_{t+1} - x_*\|_2^2\right) \end{aligned}$$

bzw.

$$\underbrace{\frac{1}{T}\sum_{t=0}^{T-1}(f_t - f_*)}_{\text{"mittlere Abweichung von } f_*\text{"}} \leq \frac{\gamma}{2}\underbrace{\frac{\sum_{t=0}^{T-1}\|f_t'\|_2^2}{T}}_{\text{"mittlerer Gradient"}} + \frac{1}{2\gamma}\underbrace{\frac{\|x_0 - x_*\|_2^2 - \|x_T - x_*\|_2^2}{T}}_{\leq \frac{\|x_0 - x_*\|_2^2}{T} = \mathcal{O}(\frac{1}{T})\text{ "Startfehler"}}.$$

Mit $\hat{t} = \operatorname{argmin}_{t\in\{0,\dots,T-1\}}(f_t - f_*)$ ($\hat{t}$ nicht notwendig gleich $T-1$) folgt

$$f_{\hat{t}} - f_* \leq \frac{1}{T}\sum_{t=0}^{T-1}(f_t - f_*).$$

Der Anteil

$$\frac{1}{2\gamma}\frac{\|x_0 - x_*\|_2^2 - \|x_T - x_*\|_2^2}{T}$$

war zu erwarten. Das Ziel ist es nun

$$\frac{\gamma}{2T}\sum_{t=0}^{T-1}\|f_t'\|_2^2$$

zu kontrollieren. Dazu muss f zusätzliche Voraussetzungen erfüllen.

1.3 Lipschitz-Stetigkeit

Im letzten Abschnitt haben wir $f \in C^1(\mathbb{R}^d)$ konvex vorausgesetzt. Jetzt fordern wir zusätzlich Lipschitz-Stetigkeit von f, d.h.

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

Dies ist äquivalent zur Beschränktheit des Gradienten, wie das folgende Ergebnis aus der Analysis zeigt.

Lemma: $f : \mathbb{R}^d \supset \operatorname{dom}(f) \to \mathbb{R}$ differenzierbar (nicht notwendig konvex), $X \subset \operatorname{dom}(f)$ offen, konvex. Dann ist

$$|f(x) - f(y)| \leq L_f\|x - y\| \quad \forall x, y \in X$$

äquivalent zu

$$\|f'(x)\| \leq L_f \quad \forall x \in X,$$

wobei bei f' die induzierte Operatornorm benutzt wird.

Beweis:

"$\Rightarrow$"

- für f gelte
$$|f(x) - f(y)| \leq L_f \|x - y\| \quad \forall x, y \in X$$
- da X offen ist gibt es für jedes $x \in X$ eine Kugel $B_r(x)$ mit $B_r(x) \subset X$
- für beliebiges $v \in \mathbb{R}^d$ mit $\|v\| = 1$ ist deshalb die Funktion
$$g(t) = f(x + tv), \quad t \in (-r, r)$$
wohldefiniert
- mit f ist auch g differenzierbar mit
$$g'(t) = f'(x + tv)v$$
und somit gilt für alle $v \in \mathbb{R}^d$ mit $\|v\| = 1$
$$\begin{aligned} \|f'(x)v\| &= |g'(0)| \\ &= \Big|\lim_{t\to 0} \frac{g(t) - g(0)}{t}\Big| \\ &= \lim_{t\to 0} \Big|\frac{f(x + tv) - f(x)}{t}\Big| \\ &\leq L_f \lim_{t\to 0} \Big\|\frac{x + tv - x}{t}\Big\| \\ &= L_f \|v\|, \end{aligned}$$
also
$$\|f'(x)\| \leq L_f$$

“⇐”

- für f gelte
$$\|f'(x)\| \leq L_f \quad \forall x \in X$$
- da X konvex ist, ist für alle $x, y \in X$ und $t \in [0, 1]$ die Funktion
$$g(t) = f\big(x + t(y - x)\big)$$
wohldefiniert und es gilt
$$g(0) = f(x), \quad g(1) = f(y)$$
- mit f ist auch g differenzierbar mit
$$g'(t) = f'\big(x + t(y - x)\big)(y - x)$$
- durch Anwendung des Mittelwertsatzes folgt
$$\begin{aligned} |f(x) - f(y)| &= |g(1) - g(0)| \\ &= |g'(\tau)| \\ &= |f'\big(\underbrace{x + \tau(y - x)}_{\xi}\big)(y - x)| \\ &\leq \|f'(\xi)\| \, \|y - x\| \\ &\leq L_f \|x - y\| \end{aligned}$$

□

Setzen wir dies in die summierte Abschätzung von oben ein, so erhalten wir

$$\begin{aligned}\sum_{t=0}^{T-1}(f_t - f_*) &\le \sum_{t=0}^{T-1} f_t'(x_t - x_*)\\ &= \frac{1}{2\gamma}\sum_{t=0}^{T-1}\left(\gamma^2\|f_t'\|_2^2 + \|x_0 - x_*\|_2^2 - \|x_{t+1} - x_*\|_2^2\right)\\ &\le \frac{\gamma}{2}TL_f^2 + \frac{1}{2\gamma}\Big(\underbrace{\|x_0 - x_*\|_2^2}_{e_0^2} - \underbrace{\|x_T - x_*\|_2^2}_{\ge 0}\Big)\\ &\le \frac{\gamma T L_f^2}{2} + \frac{e_0^2}{2\gamma},\end{aligned}$$

also

$$\min_{t=0,\dots,T-1}(f_t - f_*) \le \frac{1}{T}\sum_{t=0}^{T-1}(f_t - f_*) \le \frac{\gamma L_f^2}{2} + \frac{e_0^2}{2\gamma T}.$$

Wann verschwindet die rechte Seite für $T \to \infty$? Beide Summanden auf der rechten Seite sind ≥ 0, so dass

$$\frac{\gamma L_f^2}{2} \xrightarrow{T\to\infty} 0, \quad \frac{e_0^2}{2\gamma T} \xrightarrow{T\to\infty} 0$$

gelten muss, also

$$\gamma \xrightarrow{T\to\infty} 0, \quad \gamma T \xrightarrow{T\to\infty} \infty.$$

Mit dem Ansatz

$$\gamma = \frac{c}{T^\omega}, \quad c, \omega > 0$$

gilt immer $\gamma \xrightarrow{T\to\infty} 0$.

Für den zweiten Teil erhalten wir $\gamma T = cT^{1-\omega} \xrightarrow{T\to\infty} \infty$ falls $1 - \omega > 0$, also

$$\omega < 1$$

ist. Oben eingesetzt folgt

$$\begin{aligned}\min_{t=0,\dots,T-1}(f_t - f_*) &\le \frac{\gamma L_f^2}{2} + \frac{e_0^2}{2\gamma T}\\ &= \frac{cL_f^2}{2}\frac{1}{T^\omega} + \frac{e_0^2}{2c}\frac{1}{T^{1-\omega}}\\ &= \mathcal{O}\Big(\big(\frac{1}{T}\big)^{\min(\omega, 1-\omega)}\Big).\end{aligned}$$

Die obere Schranke

$$g(\gamma) = \frac{\gamma L_f^2}{2} + \frac{e_0^2}{2\gamma T}$$

wird wegen

$$g'(\gamma) = \frac{L_f^2}{2} - \frac{e_0^2}{2\gamma^2 T}, \quad g''(\gamma) = \frac{e_0^2}{\gamma^3 T} \ge 0$$

minimal für

$$\gamma_{\min} = \frac{e_0}{L_f\sqrt{T}}$$

mit

$$g_{\min} = g(\gamma_{\min}) = \frac{L_f e_0}{\sqrt{T}}.$$

Damit erhalten wir das folgende Ergebnis.

Satz: $f : \mathbb{R}^d \to \mathbb{R}$, konvex, C^1, 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

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

Die optimale Ordnung ist $\frac{1}{2}$ bei $\omega = \frac{1}{2}$.

Mit $e_0 = \|x_0 - x_*\|_2$, $\gamma = \frac{e_0}{L_f\sqrt{T}}$ gilt außerdem

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

Bemerkung:

- $\min_{t=0,\dots,T-1}(f_t - f_*) \le \varepsilon$ gilt damit sicher, falls

  $$\frac{L_f e_0}{\sqrt{T}} \le \varepsilon$$

  bzw.

  $$T \ge \big(\frac{e_0}{L_f\varepsilon}\big)^2$$

- für $\min_{t=0,\dots,T-1}(f_t - f_*) \le \varepsilon$ benötigen wir damit höchstens $\mathcal{O}(\frac{1}{\varepsilon^2})$ Schritte
- in der Praxis gibt man ε vor, bestimmt T und das zugehörige (feste)

  $$\gamma = \frac{e_0}{L_f\sqrt{T}}$$

  und führt dann (maximal) $T-1$ Schritte des Verfahrens durch
- für $\varepsilon \to 0$ gilt $T\to\infty$ und

  $$\gamma = \frac{e_0}{L_f\sqrt{T}} \to 0$$

1.4 L-Glattheit

Ist f konvex und C^1, dann gilt

$$f(y) \ge f(x) + f'(x)(y-x),$$

d.h. der Graph von f verläuft oberhalb seiner Tangenten. Zur Abschätzung nach oben führen wir den folgenden Begriff ein.

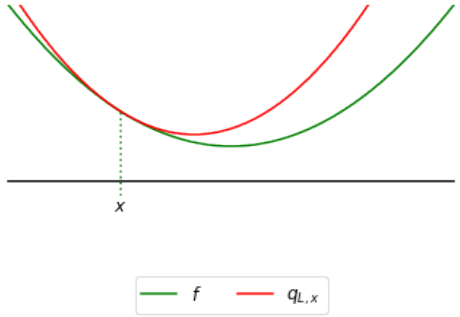
Definition: $f : \mathbb{R}^d \supset \operatorname{dom}(f) \to \mathbb{R}$ (nicht notwendig konvex), $X \subset \operatorname{dom}(f)$. f heißt L-glatt auf X falls ein $L > 0$ existiert, mit

$$f(y) \le f(x) + f'(x)(y-x) + \frac{1}{2}L\|y-x\|_2^2 \quad \forall x,y \in X.$$

Bemerkung: Ist f L-glatt, so verläuft der Graph von f unterhalb der quadratischen Approximation

$$q_{L,x}(y) = f(x) + f'(x)(y-x) + \frac{1}{2}L\|y-x\|_2^2.$$

```
import sympy as sy
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline

fsize = 12

x = sy.symbols('x')
f  = sy.Lambda(x, x*x/2 + 1/2)
f1 = sy.Lambda(x, f(x).diff(x))

L  = 3/4
x0 = -1

ql = sy.Lambda(x, f(x0) + f1(x0)*(x-x0) + L*(x-x0)**2)

ql = sy.lambdify(x, ql(x))
f  = sy.lambdify(x, f(x))

x = np.linspace(-2,2)
xmin = x.min()
xmax = x.max()
plt.plot(x, f(x) , 'g', label = '$f$')
plt.plot(x, ql(x), 'r', label = '$q_{L,x}$')
plt.axis('off')
#
tic = 0.2
plt.plot([xmin, xmax], [0,0], 'k')
#
plt.text(x0, -tic, '$x$', ha = 'center', va = 'top', fontsize=fsize)
plt.plot([x0,x0], [-tic,f(x0)], 'g:')
#
plt.legend(loc='lower center', ncol=3, fontsize=fsize)
plt.ylim(xmin, f(xmin));
```



L-Glattheit ist eng verknüpft mit der Lipschitz-Stetigkeit des Gradienten f'.

Lemma: $f : \mathbb{R}^d \supset \mathrm{dom}(f) \to \mathbb{R}$ sei differenzierbar (nicht notwendig konvex). Ist f' Lipschitz-stetig, d.h.

$$\|f'(y) - f'(x)\| \le L\|y - x\| \quad \forall x, y,$$

dann gilt

$$\big(f'(y) - f'(x)\big)(y - x) \le L\|y - x\|^2.$$

Beweis: Da $f'(x), f'(y)$ lineare stetige Operatoren sind gilt

$$\begin{aligned}\big(f'(y) - f'(x)\big)(y - x) &\le \big|\big(f'(y) - f'(x)\big)(y - x)\big| \\ &\le \|f'(y) - f'(x)\|\,\|y - x\| \\ &\le L\|y - x\|^2\end{aligned}$$

□

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

$$\big(f'(y) - f'(x)\big)(y - x) \leq L\|y - x\|^2 \quad \forall x, y$$

äquivalent zu

$$f(y) \leq f(x) + f'(x)(y - x) + \frac{1}{2}L\|y - x\|^2 \quad \forall x, y.$$

Beweis:

"⇒"

- mit

  $$g(t) = f\big(x + t(y - x)\big)$$

  folgt

  $$g'(t) = f'\big(x + t(y - x)\big)(y - x)$$

  und für $t > 0$

  $$\begin{aligned} g'(t) - g'(0) &= \big(f'\big(x + t(y - x)\big) - f'(x)\big)(y - x) \\ &= \frac{1}{t}\big(f'\big(x + t(y - x)\big) - f'(x)\big)t(y - x) \\ &\leq \frac{1}{t}L\|t(y - x)\|^2 \\ &= tL\|y - x\|^2 \end{aligned}$$

- damit erhalten wir

  $$\begin{aligned} f(y) &= g(1) \\ &= g(0) + \int_0^1 g'(\tau)\, d\tau \\ &\leq f(x) + \int_0^1 g'(0) + \tau L\|y - x\|^2\, d\tau \\ &= f(x) + f'(x)(y - x) + \frac{1}{2}L\|y - x\|^2 \end{aligned}$$

"⇐"

- nach Voraussetzung ist

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

  bzw.

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

- Addition der beiden Ungleichungen liefert

  $$f(y) + f(x) \leq f(x) + f(y) + \big(f'(x) - f'(y)\big)(y - x) + L\|y - x\|^2,$$

  also

  $$\big(f'(y) - f'(x)\big)(y - x) \leq L\|y - x\|^2$$

□

Bemerkung: Ist f' Lipschitz-stetig, dann ist f L-glatt.

Ist f konvex, $\mathrm{dom}(f) = \mathbb{R}^d$ und existiert ein $x_* \in \mathrm{dom}(f)$ mit $f(x_*) = \inf_x f(x)$, dann gilt auch die Umkehrung.

Lemma: $f : \mathbb{R}^d \to \mathbb{R}$ sei differenzierbar (nicht notwendig konvex) und es existiere x_* mit $f(x_*) = \inf_{x \in \mathbb{R}^d} f(x)$. Ist f L-glatt mit Konstante L, dann gilt

$$\frac{1}{2L}\|f'(x)\|_2^2 \le f(x) - f(x_*) \le \frac{L}{2}\|x - x_*\|_2^2$$

und f' ist Lipschitz-stetig mit Konstante L.

Beweis:

- Abschätzung nach oben:
  - da x_* globaler Minimierer ist muss $f'(x_*) = 0$ sein und somit

$$\begin{aligned} f(x) &\le f(x_*) + f'(x_*)(x - x_*) + \frac{1}{2}L\|x - x_*\|_2^2 \\ &= f(x_*) + \frac{1}{2}L\|x - x_*\|_2^2 \end{aligned}$$

- Abschätzung nach unten:
  - wir benutzen

$$\begin{aligned} f(x_*) &= \inf_y f(y) \le \inf_y u(y), \\ u(y) &= f(x) + f'(x)(y - x) + \frac{1}{2}L\|y - x\|_2^2 \end{aligned}$$

    und minimieren die quadratische Funktion u
  - für die Ableitungen erhalten wir

$$u'(y) = f'(x) + L(y - x), \quad u''(y) = LI$$

  - u ist (strikt) konvex mit globalem Minimierer y_* mit

$$0 = u'(y_*) \quad \Leftrightarrow \quad y_* - x = -\frac{1}{L}f'(x)$$

    und Minimum

$$\begin{aligned} u_* = u(y_*) &= f(x) + f'(x)(y_* - x) + \frac{1}{2}L\|y_* - x\|_2^2 \\ &= f(x) - \frac{1}{L}\|f'(x)\|_2^2 + \frac{1}{2}L\|\frac{1}{L}f'(x)\|_2^2 \\ &= f(x) - \frac{1}{2L}\|f'(x)\|_2^2 \end{aligned}$$

  - somit folgt

$$f(x_*) \le f(x) - \frac{1}{2L}\|f'(x)\|_2^2$$

- Lipschitz-Stetigkeit von f':
  - wir betrachten die Funktion

$$g(y) = f(y) - f'(x)y$$

  - mit f ist auch g konvex und differenzierbar mit Ableitung

$$g'(y) = f'(y) - f'(x)$$

  - damit ist $g'(x) = 0$, also ist $y_* = x$ globales Minimum von g

- außerdem folgt aus der L-Glattheit von f für beliebiges z

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

  so dass auch g L-glatt ist

- somit können wir die Abschätzung nach unten aus dem vorherigen Teil auf g anwenden und erhalten wegen $y_* = x$

$$\begin{aligned} \frac{1}{2L}\|f'(y) - f'(x)\|_2^2 = \frac{1}{2L}\|g'(y)\|_2^2 \\ &\leq g(y) - g(y_*) \\ &= g(y) - g(x) \\ &= f(y) - f'(x)y - \big(f(x) - f'(x)x\big) \\ &= f(y) - f(x) - f'(x)(y-x), \end{aligned}$$

  also

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

  bzw. durch vertauschen von x und y

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

- durch Addition der beiden Ungleichungen erhalten wir

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

  und mit Cauchy-Schwartz

$$\begin{aligned} \|f'(y) - f'(x)\|_2^2 \leq L\big(f'(y) - f'(x)\big)(x-y) \\ &\leq L\|f'(y) - f'(x)\|_2\|x-y\|_2, \end{aligned}$$

  so dass f' Lipschitz-stetig mit Konstante L ist

□

Bemerkung: Sei $f : \mathbb{R}^d \to \mathbb{R}$ konvex, C^1 und es existiere $x_* \in \mathrm{dom}(f)$ mit $f(x_*) = \inf_x f(x)$. Dann ist äquivalent:

- f ist L-glatt mit Parameter L
- $\|f'(y) - f'(x)\|_2 \leq L\|y-x\|_2 \quad \forall x, y \in \mathbb{R}^d$

L-Glattheit ist also unter diesen Voraussetzungen äquivalent dazu, dass f' Lipschitz-stetig mit Konstante L ist.

Folgende Operation erhalten die L-Glattheit:

- für $i = 1, \ldots, m$ seien $f_i : \mathbb{R}^d \supset \mathrm{dom}(f_i) \to \mathbb{R}$, L-glatt mit Parameter L_i und $\lambda_i \geq 0$. Dann ist

$$f = \sum_{i=1}^{m} \lambda_i f_i$$

L-glatt mit Konstante

$$L = \sum_{i=1}^{m} \lambda_i L_i$$

über

$$\operatorname{dom}(f) = \bigcap_{i=1}^{m} \operatorname{dom}(f_i)$$

- ist $f : \mathbb{R}^d \supset \operatorname{dom}(f) \to \mathbb{R}$ L-glatt mit Konstante L, $g : \mathbb{R}^m \to \mathbb{R}^d$ affin linear, d.h.

$$g(z) = Az + b,$$

dann ist $f \circ g$ auch L-glatt mit

$$\tilde{L} = L\,\|A\|_2^2, \quad \operatorname{dom}(f \circ g) = \{z \,|z \in \mathbb{R}^m,\ g(z) \in \operatorname{dom}(f)\}$$

Für f konvex und L-glatt werden wir nun günstigere Konvergenzresultate für Gradient-Descent erhalten. Wir benutzen L-Glattheit mit $y = x_{t+1}$, $x = x_t$

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

Mit $x_{t+1} - x_t = -\gamma f_t$ folgt

$$f_{t+1} \le f_t - \gamma\|f_t'\|_2^2 + \frac{L}{2}\gamma^2\|f_t'\|_2^2 = f_t - \underbrace{\gamma\big(1 - \frac{L}{2}\gamma\big)}_{=:\beta}\|f_t'\|_2^2.$$

Für $\gamma > 0$ ist $\beta > 0$ genau dann, wenn

$$1 - \frac{L}{2}\gamma > 0,$$

also genau dann, wenn

$$0 < \gamma < \frac{2}{L}.$$

Damit folgt

Descent-Lemma: Ist $f : \mathbb{R}^d \to \mathbb{R}$ L-glatt, dann gilt

$$f_{t+1} \le f_t - \beta\|f_t'\|_2^2, \quad \beta = \gamma\big(1 - \frac{\gamma L}{2}\big)$$

und $\beta > 0$ falls $0 < \gamma < \frac{2}{L}$.

Bemerkung: Für f L-glatt und $0 < \gamma < \frac{2}{L}$ fällt f_t also *monoton*.

Damit verschärfen wir jetzt unser Konvergenzresultat aus dem vorherigen Abschnitt. Nach den Vorüberlegungen gilt

$$\sum_{t=0}^{T-1}(f_t - f_*) \le \frac{\gamma}{2}\sum_{t=0}^{T-1}\|f_t'\|_2^2 + \frac{1}{2\gamma}\|x_0 - x_*\|_2^2 - \frac{1}{2\gamma}\|x_T - x_*\|_2^2.$$

Aus dem Descent-Lemma folgt

$$\|f_t'\|_2^2 \le \frac{1}{\beta}(f_t - f_{t+1})$$

und somit

$$\begin{aligned}\sum_{t=0}^{T-1}(f_t - f_*) &\le \frac{\gamma}{2\beta}\sum_{t=0}^{T-1}(f_t - f_{t+1}) + \frac{1}{2\gamma}\big(\|x_0 - x_*\|_2^2 - \|x_T - x_*\|_2^2\big)\\ &= \frac{\gamma}{2\beta}(f_0 - f_T) + \frac{1}{2\gamma}\big(\|x_0 - x_*\|_2^2 - \|x_T - x_*\|_2^2\big).\end{aligned}$$

Für $\gamma < \frac{2}{L}$ ist $f_{t+1} \leq f_t$, so dass

$$f_{T-1} \leq \frac{1}{T} \sum_{t=0}^{T-1} f_t$$

und damit

$$\begin{aligned} f_{T-1} - f_* &\leq \frac{1}{T} \sum_{t=0}^{T-1} (f_t - f_*) \\ &\leq \frac{1}{T}\Big(\frac{\gamma}{2\beta}(f_0 - f_T) + \frac{1}{2\gamma}\big(\|x_0 - x_*\|_2^2 - \|x_T - x_*\|_2^2 \big) \Big). \end{aligned}$$

Mit $\|x_T - x_*\|_2 \geq 0$ erhalten wir schließlich

$$f_{T-1} - f_* \leq \frac{1}{T}\Big(\frac{\gamma}{2\beta}(f_0 - f_*) + \frac{1}{2\gamma}\|x_0 - x_*\|_2^2 \Big).$$

Satz: $f : \mathbb{R}^d \to \mathbb{R}$, konvex, L-glatt mit Konstante L und es existiere $x_* = \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$. Für $0 < \gamma < \frac{2}{L}$ ist

$$\begin{aligned} f_T - f_* &\leq \frac{1}{T+1}\Big(\frac{\gamma}{2\beta}(f_0 - f_*) + \frac{1}{2\gamma}\|x_0 - x_*\|_2^2 \Big) \\ &= \mathcal{O}\big(\frac{1}{T}\big) \quad \text{für} \quad T \to \infty. \end{aligned}$$

Bemerkung:

- $f_T - f_* \leq \varepsilon$ gilt damit sicher, falls

  $$\frac{1}{T+1}\Big(\frac{\gamma}{2\beta}(f_0 - f_*) + \frac{1}{2\gamma}\|x_0 - x_*\|_2^2 \Big) \leq \varepsilon$$

  also

  $$T \geq \frac{1}{\varepsilon}\Big(\frac{\gamma}{2\beta}(f_0 - f_*) + \frac{1}{2\gamma}\|x_0 - x_*\|_2^2 \Big) - 1 = \mathcal{O}\big(\frac{1}{\varepsilon}\big)$$

- da $\gamma < \frac{2}{L}$ sein muss kann wegen des Terms

  $$\frac{1}{2\gamma}\|x_0 - x_*\|_2^2$$

  die Asymptotik der oberen Schranke nicht mehr durch eine T-abhängige Wahl von γ verbessert werden

1.5 μ-Konvexität

Bis jetzt haben wir nur Abschätzungen für $f_t - f_*$ bewiesen. Nun werden wir $\|x_t - x_*\|_2$ betrachten. Dazu benötigen wir nochmals stärkere Voraussetzungen, nämlich μ-Konvexität. Damit werden wir zusätzlich auch eine besseres asymptotisches Verhalten nachweisen können.

f ist μ-konvex falls $f(x) - \frac{\mu}{2}\|x\|_2^2$ konvex ist. Ist f zusätzlich differenzierbar, so erhalten wir das folgende Ergebnis.

Lemma: Ist f μ-konvex und differenzierbar, dann gilt

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

und

$$\big(f'(y) - f'(x)\big)(y - x) \geq \mu\|y - x\|_2^2 \quad \forall x, y.$$

Beweis:

- f ist μ-konvex falls $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ konvex ist

- mit f ist auch g differenzierbar mit
$$g'(x) = f'(x) - \mu x$$
- wegen
$$g(y) \geq g(x) + g'(x)(y-x)$$
ist
$$f(y) - \frac{\mu}{2}\|y\|_2^2 \geq f(x) - \frac{\mu}{2}\|x\|_2^2 + \big(f'(x) - \mu x\big)^T (y-x)$$
bzw.
$$f(y) \geq f(x) + f'(x)(y-x) + \frac{\mu}{2}\Big(\underbrace{y^T y - x^T x - 2x^T(y-x)}_{h}\Big)$$
mit
$$\begin{aligned} h &= y^T y - x^T x - 2x^T y + 2x^T x \\ &= y^T y - 2x^T y + x^T x \\ &= \|y-x\|_2^2, \end{aligned}$$
also
$$f(y) \geq f(x) + f'(x)(y-x) + \frac{\mu}{2}\|y-x\|_2^2$$
da g konvex und differenzierbar ist, ist g' monoton, also
$$\begin{aligned} 0 &\leq \big(g'(y) - g'(x)\big)(y-x) \\ &= \big(f'(y) - \mu y - f'(x) + \mu x\big)(y-x) \\ &= \big(f'(y) - f'(x)\big)(y-x) - \mu\|y-x\|_2^2 \end{aligned}$$
und somit
$$\big(f'(y) - f'(x)\big)(y-x) \geq \mu\|y-x\|_2^2 \quad \forall x, y$$

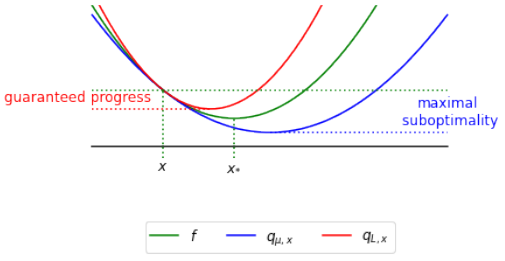
□

Bemerkung:

- ist f μ-konvex und L-glatt (und damit differenzierbar), so gilt
$$\begin{aligned} f(y) &\leq f(x) + f'(x)(y-x) + \frac{L}{2}\|y-x\|_2^2 =: q_{L,x}(y) \\ f(y) &\geq f(x) + f'(x)(y-x) + \frac{\mu}{2}\|y-x\|_2^2 =: q_{\mu,x}(y) \end{aligned}$$
- f kann also zwischen den beiden quadratischen Funktionen $q_{\mu,x}$, $q_{L,x}$ "eingesperrt" werden
- $q_{\mu,x}$, $q_{L,x}$ berühren f im Punkt x
- es muss immer $\mu \leq L$ gelten
- ist $\mu > 0$ so ist f strikt konvex und x_* ist damit eindeutig
- ist $\mu = L$ dann ist $f = q_{\mu,x} = q_{L,x}$, d.h. f ist ein quadratisches Polynom

```
import sympy as sy
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline

fsize = 12

x = sy.symbols('x')
```

```
f  = sy.Lambda(x, x*x/2 + 1/2)
f1 = sy.Lambda(x, f(x).diff(x))

m  = 1/3
L  = 3/4
x0 = -1

qm = sy.Lambda(x, f(x0) + f1(x0)*(x-x0) + m*(x-x0)**2)
ql = sy.Lambda(x, f(x0) + f1(x0)*(x-x0) + L*(x-x0)**2)


x1 = sy.solve(ql(x).diff(x))[0]
x2 = sy.solve(f1(x))[0]
x3 = sy.solve(qm(x).diff(x))[0]

qm = sy.lambdify(x, qm(x))
ql = sy.lambdify(x, ql(x))
f  = sy.lambdify(x, f(x))

x = np.linspace(-2,3)
xmin = x.min()
xmax = x.max()
plt.plot(x, f(x) , 'g', label = '$f$')
plt.plot(x, qm(x), 'b', label = '$q_{\mu,x}$')
plt.plot(x, ql(x), 'r', label = '$q_{L,x}$')
plt.axis('off')
#
tic = 0.2
plt.plot([xmin, xmax], [0,0], 'k')
#
plt.text(x2, -tic, '$x_{*}$', ha = 'center', va = 'top', fontsize=fsize)
plt.plot([x2,x2], [-tic,f(x2)], 'g:')
#
plt.text(x0, -tic, '$x$', ha = 'center', va = 'top', fontsize=fsize)
plt.plot([x0,x0], [-tic,f(x0)], 'g:')
plt.plot([xmin,x0], [f(x0),f(x0)], 'g:')
#
plt.plot([xmin,x1], [ql(x1),ql(x1)], 'r:')
#
plt.text(xmin-tic, (f(x0)+f(x1)+tic)/2, 'guaranteed progress', color='r', ha =␣
 ↪'center', va = 'center', fontsize=fsize)
#
#
plt.plot([x0,xmax], [f(x0),f(x0)], 'g:')
plt.plot([x3,xmax], [qm(x3),qm(x3)], 'b:')
plt.text(xmax, (f(x0)+qm(x3))/2, 'maximal\n suboptimality', color='b', ha = 'center',␣
 ↪va = 'center', fontsize=fsize)
#
plt.legend(loc='lower center', ncol=3, fontsize=fsize)
plt.ylim(xmin, f(xmin));
```



Lemma: $f : \mathbb{R}^d \to \mathbb{R}$ differenzierbar und es existiere x_* mit $f(x_*) = \inf_{x\in\mathbb{R}^d} f(x)$. Ist f μ-konvex dann gilt

$$\frac{\mu}{2}\|y-x\|_2^2 \le f(x) - f(x_*) \le \frac{1}{2\mu}\|f'(x)\|_2^2.$$

Beweis:

- f ist μ-konvex, d.h.
$$f(y) \geq f(x) + f'(x)(y-x) + \frac{\mu}{2}\|y-x\|_2^2 \quad \forall x, y$$
- Abschätzung nach unten:
  - da x_* globaler Minimierer ist muss $f'(x_*) = 0$ sein und aus der μ-Konvexität folgt
$$\begin{aligned} f(x) &\geq f(x_*) + f'(x_*)(x-x_*) + \frac{\mu}{2}\|x-x_*\|_2^2 \\ &= f(x_*) + \frac{\mu}{2}\|x-x_*\|_2^2 \end{aligned}$$
- Abschätzung nach oben:
  - wir benutzen
$$\begin{aligned} f(x_*) &= \inf_y f(y) \geq \inf_y u(y), \\ u(y) &= f(x) + f'(x)(y-x) + \frac{\mu}{2}\|y-x\|_2^2 \end{aligned}$$
    und minimieren die quadratische Funktion u
  - für die Ableitungen erhalten wir
$$u'(y) = f'(x) + \mu(y-x), \quad u''(y) = \mu I$$
  - für $\mu > 0$ ist u strikt konvex mit globalem Minimierer y_* mit
$$0 = u'(y_*) \quad \Leftrightarrow \quad y_* - x = -\frac{1}{\mu} f'(x)$$
    und Minimum
$$\begin{aligned} u_* = u(y_*) &= f(x) + f'(x)(y_*-x) + \frac{1}{2}\mu\|y_*-x\|_2^2 \\ &= f(x) - \frac{1}{\mu}\|f'(x)\|_2^2 + \frac{1}{2}\mu\|\frac{1}{\mu}f'(x)\|_2^2 \\ &= f(x) - \frac{1}{2\mu}\|f'(x)\|_2^2 \end{aligned}$$
  - somit folgt
$$f(x_*) \geq f(x) - \frac{1}{2\mu}\|f'(x)\|_2^2.$$
    bzw.
$$f(x) - f(x_*) \leq \frac{1}{2\mu}\|f'(x)\|_2^2$$

□

Aus $x_{t+1} = x_t - \gamma f_t'$ hatten wir in den Vorüberlegungen

$$\|x_{t+1} - x_*\|_2^2 = \|x_t - x_*\|_2^2 + \gamma^2\|f_t'\|_2^2 - 2\gamma f_t'(x_t - x_*)$$

erhalten. μ-Konvexität liefert mit $y = x_*$, $x = x_t$

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

bzw.

$$-f_t'(x_t - x_*) \le f_* - f_t - \frac{\mu}{2}\|x_t - x_*\|_2^2.$$

Eingesetzt erhalten wir

$$\begin{aligned}\|x_{t+1} - x_*\|_2^2 &= \|x_t - x_*\|_2^2 + \gamma^2\|f_t'\|_2^2 + 2\gamma\big(f_* - f_t - \frac{\mu}{2}\|x_t - x_*\|_2^2\big)\\ &\le (1 - \gamma\mu)\|x_t - x_*\|_2^2 + \gamma^2\|f_t'\|_2^2 + 2\gamma(f_* - f_t).\end{aligned}$$

Das Descent-Lemma aus dem vorherigen Kapitel liefert

$$f_* - f_t \le f_{t+1} - f_t \le -\beta\|f_t'\|_2^2, \quad \beta = \gamma\big(1 - \frac{\gamma L}{2}\big)$$

mit $\beta > 0$ für $0 < \gamma < \frac{2}{L}$, so dass

$$\|x_{t+1} - x_*\|_2^2 \le (1 - \gamma\mu)\ \|x_t - x_*\|_2^2 + (\gamma^2 - 2\gamma\beta)\ \|f_t'\|_2^2$$

gilt.

Für den Vorfaktor des letzten Terms gilt wegen $\gamma > 0$, $\beta = \gamma\big(1 - \frac{\gamma L}{2}\big)$,

$$\begin{aligned}\gamma(\gamma - 2\beta) \le 0 \quad &\Leftrightarrow \quad \gamma \le 2\beta = \gamma(2 - \gamma L)\\ &\Leftrightarrow \quad 1 \le 2 - \gamma L\\ &\Leftrightarrow \quad \gamma \le \frac{1}{L}.\end{aligned}$$

Somit folgt für $0 < \gamma \le \frac{1}{L}$

$$\|x_{t+1} - x_*\|_2^2 \le \rho\|x_t - x_*\|_2^2, \quad \rho = 1 - \gamma\mu$$

bzw.

$$\|x_T - x_*\|_2^2 \le \rho^T\|x_0 - x_*\|_2^2.$$

Wegen $0 < \mu \le L$ und $0 < \gamma \le \frac{1}{L}$ ist

$$0 < \gamma\mu \le \gamma L \le 1$$

und somit

$$0 \le \rho < 1,$$

also $|\rho| < 1$ und wir erhalten Konvergenz für x_T.

Für f_T ergibt sich direkt aus der L-Glattheit

$$f_T \le f_* + f_*'(x_T - x_*) + \frac{L}{2}\|x_T - x_*\|_2^2$$

und wegen $f_*' = 0$

$$f_T - f_* \le \frac{L}{2}\|x_T - x_*\|_2^2 \le \frac{L}{2}\rho^T\|x_0 - x_*\|_2^2.$$

Insgesamt haben wir damit das folgende Ergebnis bewiesen.

Satz: $f : \mathbb{R}^d \to \mathbb{R}$, μ-konvex mit $\mu > 0$, L-glatt mit Konstante L und es existiere $x_* = \operatorname{argmin}_{x\in\mathbb{R}^d} f(x)$.

Für $0 < \gamma \le \frac{1}{L}$ folgt

$$\|x_{t+1} - x_*\|_2^2 \le \rho\|x_t - x_*\|_2^2$$

und

$$f_T - f_* \le \frac{L}{2}\rho^T\|x_0 - x_*\|_2^2$$

mit

$$\rho = 1 - \gamma\mu \in [0, 1).$$

Bemerkung:

- $f_T - f_* \leq \frac{L}{2}\rho^T \|x_0 - x_*\|_2^2 \leq \varepsilon$ gilt sicher, falls

$$\rho^T \leq \frac{2\varepsilon}{L\|x_0 - x_*\|_2^2}$$

- für $\rho = 0$ gilt das für alle $\varepsilon \geq 0$
- für $0 < \rho < 1$ folgt

$$T\log(\rho) \leq \log\Big(\frac{2\varepsilon}{L\|x_0 - x_*\|_2^2}\Big), \quad \log(\rho) < 0$$

  also

$$\begin{aligned} T &\geq \frac{1}{\log(\rho)} \log\Big(\frac{2\varepsilon}{L\|x_0 - x_*\|_2^2}\Big) \\ &= \frac{1}{|\log(\rho)|} \log\Big(\frac{L\|x_0 - x_*\|_2^2}{2\varepsilon}\Big) \\ &= \frac{1}{|\log(\rho)|}\Big(\log\Big(\frac{L}{2}\|x_0 - x_*\|_2^2\Big) + \log\Big(\frac{1}{\varepsilon}\Big)\Big) \end{aligned}$$

  und somit

$$T = \mathcal{O}\Big(\log\Big(\frac{1}{\varepsilon}\Big)\Big)$$

1.6 Zusammenfassung

Für Gradient-Descent bei *nicht restringierten Optimierungsproblemen* haben wir folgendes Konvergenzverhalten nachgewiesen:

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

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

- f konvex und L-glatt, $0 < \gamma < \frac{2}{L}$:

$$f_T - f_* \leq \varepsilon \quad \Rightarrow \quad T = \mathcal{O}\Big(\frac{1}{\varepsilon}\Big)$$

- f μ-konvex mit $\mu > 0$ und L-glatt, $0 < \gamma \leq \frac{1}{L}$:

$$f_T - f_* \leq \varepsilon \quad \Rightarrow \quad T = \mathcal{O}\Big(\log\Big(\frac{1}{\varepsilon}\Big)\Big)$$

Ist $f(x)$ eine L-glatte Funktion, dann ist für alle $\mu > 0$

$$f_R(x) = f(x) + \mu\|x\|_2^2$$

auch μ-konvex, d.h. Tikhonov(Ridge)-Regularisierung kann die Konvergenz von Gradient-Descent beschleunigen.