Labor 4: Optimierung

Thema: Regression

Inhalte des Matlab-Crashkurses

- find, diff, load
- symbolisch rechnen mit Matlab: syms, diff, solve
- global

Ableitungen und Glättung diskreter Funktionen

(a) Machen Sie sich mit dem Programm `MyDiskDiff.m` vertraut.

(b) Verwenden Sie statt der Beispielfunktionen Daten aus der Datei `Glocke.dat`. Bestimmen Sie Extremalstellen $E = (x_E, y_E)$ und Wendepunkte $W = (x_W, y_W)$.

(c) Wiederholen Sie Teil (b) mit den Daten `GlockeR.dat`. Das Rauschen sollte Ihnen Schwierigkeiten bereiten.

(d) Glätten Sie die Daten mit der Funktion `Smooth` im Skript `MyDataSmoother.m`

```
// y mit Rauschen
Loop  = ____;
delta = ____;

for i=1:Loop
   d2y = diskdiff(y,F,1);
   y = y + delta*d2y;
end
// y geglaettet
```

Erszeugen Sie einen Plot von den Originaldaten gegenüber den geglätteten. Spielen Sie mit `delta` und `Loop`.

Tipp: Wählen Sie zunächst `Loop=1` und für `delta` den größtmöglichen Wert, der in die richtige Richtung führt. Dann steigern Sie `Loop`.

Ihr Ergebnis könnte so aussehen wie in Abbildung 1.

(e) Widerholen Sie teil (b) mit den nun geglätteten Daten.

Parameter einer Funktion bestimmen

Wir betrachten die Ansatzfunktion (kennen Sie schon aus dem Kino-Beispiel)

$$g(x) = a\,\mathrm{e}^{-\frac{(x-b)^2}{c}}\,.$$

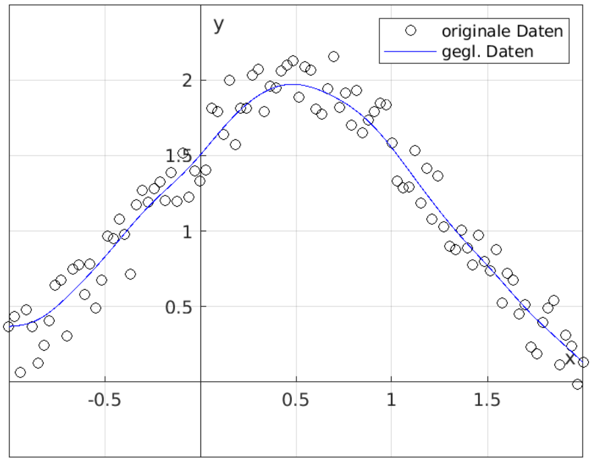

Abbildung 1: Geglättete Daten "GlockeR.dat"

(f) Berechnen Sie Bedingungen an die Parameter, so dass gewünschte Positionen von Extremwert $E = (x_E, y_E)$ und Wendepunkten $W = (x_W, y_W)$ erfüllt sind. Anmerkung: Wegen der Symmetrie braucht nur ein Wendepunkt berücksichtigt werden.

Die Ableitungen sind (können Sie sich auch von Matlab berechnen lassen)

$$g'(x) = -a\,\frac{2\,(x-b)}{c}\,\mathrm{e}^{-\frac{(x-b)^2}{c}}$$

$$g''(x) = \left(-\frac{2}{c} + \frac{4\,(x-b)^2}{c^2}\right)\,a\,\mathrm{e}^{-\frac{(x-b)^2}{c}}$$

Auch beim Extrahieren von Parametern können Sie sich mittels symbolichen Rechnens von Matlab helfen lassen:

$$g'(x_E) = 0 \quad \Leftrightarrow \quad b = x_E$$

$$g''(x_W) = 0 \quad \Leftrightarrow \quad c = 2\,(x_W - b)^2$$

$$\max_I g(x) = g(x_E) \quad \Leftrightarrow \quad a = \max_i y(i)$$

x_E, x_W und $\max_i y(i)$ kennen wir aus den Daten; Teil (c)(iii). Ihre Lösung sollte so aussehen wie in Abbildung 2 dargestellt.

(g) Plotten Sie mit den gefundenen Parametern die Funktion g zusammen mit den Daten. Wahrscheinlich sieht das schon ganz gut aus.

(h) Berechnen Sie den Fehler

$$Err = \sqrt{\mathrm{Res}_1^2 + \cdots + \mathrm{Res}_N^2}\,,$$

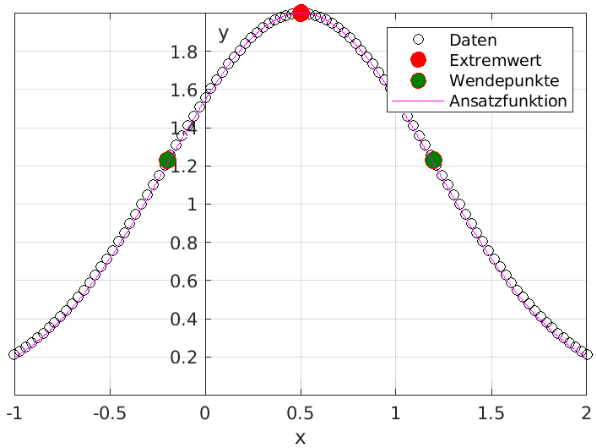

Abbildung 2: Lösung zu den Daten "Glocke.dat"

wobei das Residuum Res_i jeweils den Abstand zwischen Funktionswert $g(D_{i,x})$ und entsprechenden y-Werten der Daten $D_{i,y}$ beschreibt. Wenn sie an den Parameter a, b, c wackeln werden Sie eine Veränderung des Fehlers bemerken.

Regression mit Matlab

(i) Machen Sie sich mit dem PRogramm `MyRegression.m` vertraut.

(j) Berechnen Sie die Regression für die Daten `Glocke.dat` und `GlockeR.dat`. Verwenden Sie die grob geschätzten Parameter von oben als Startwerte für die Regression.

(k) Die Berechnung der Regression erfolgt über eine Minimierung der Fehlerfunktion Err (Quadratsumme der Residuen). Dabei wird mit einem Newton-Verfahren näherungsweise die Nullstelle der ersten Ableitung (multivariat, wenn es mehr als einen zu betimmenden Parameter gibt) berechnet. Manchmal ist es nötig, in der Nähe dieser Nullstelle zu starten. Mit den grob geschätzen Parametern haben wir gute Startwerte bereits gefunden.

Im Fall der Glockenkurve ist die Wahl von geschätzten Parametern als Startwert nicht nötig. Anders wird es sich beim Beispiel in der Aufgabenstellung bei der Aufgabe zur Abgabe verhalten!

Laboraufgabe 1: Aufgabe zur Abgabe

Die Spanische Grippe kam nicht aus Spanien. Vermutlich gelangte sie von den USA mit Truppentransporten nach Europa. Mitte April 1918 wurden die ersten Fälle in Frankreich registriert. Die Medizin stand vor einem Rätsel: Viren als Krankheitsauslöser wurden erst 1933 entdeckt.

Um die Bevölkerung nicht weiter zu beunruhigen, gab es in den kriegsführenden Staaten kaum Nachrichten darüber. Spanien jedoch war neutral. Die Zeitungen berichteten unzensiert und erweckten so den Eindruck, die Seuche habe hier ihren Ursprung. (Quelle: https://erster-weltkrieg.dnb.de)

Abbildung 3: Spanische Grippe 1918-1920

(i) Im Datensatz `influenza_england_1978_school.dat` finden Sie Daten einer Epidemie einer englischen Schule von 1978 und im Datensatz Daten zur Covid-Epidemie in Deutschland. Laden Sie einen der Datensätze nach Wahl. In der ersten Spalte befindet sich jeweils die Angabe von Tagen und in der zweiten Spalte die Angabe von neu infizierten Personen.

(ii) Glätten Sie zunächst die Daten. Nicht zu viel. Gerade so viel, dass Sie mit den geglätteten Daten qualitativ gute Ergebnisse erreichen können.

(iii) Bestimmen Sie die charakteristischen Größen (Extrema und Wendepunkte) der geglätteten Daten rechnerisch mit diskreten Ableitungen. (Versuchen Sie das auch mal ohne die Glättung davor.)

(iv) Ermitteln Sie automatisiert aus den charakteristischen Größen die Parameter α, β und γ der Ansatzfunktion.

$$f(x) = \alpha \operatorname{sech}^2(\beta\, t - \gamma)\,.$$

Der Sekanshyperbolikus ist der Kehrwert des Kosinushyperbolikus:

$$\cosh(x) = \frac{1}{2}\left(\mathrm{e}^{x} + \mathrm{e}^{-x}\right)$$

(v) Ploten Sie die Daten zusammen mit der Ansatzfunktion und Extremalstellen und Wendepunkte.

(vi) Welche Werte für Extremalstelle und Wendepunkte hätten Sie augenscheinlich erwartet?

(vii) Berechnen Sie die Regression mit der Matlabfunktion `lsqnonlin`. Verwenden Sie als Startwerte einmal Nullen und dann Ihre grobe Abschätzung der Parameter. Stellen Sie die Fehler gegenüber.

(viii) An welchem Tag gibt es wahrscheinlich keine infizierten Personen mehr? Was ist Ihr Entscheidungskriterium?