

lab_nlls_partial

November 8, 2022

1 Lab: Nonlinear Least Squares for Modeling Materials

Nonlinear least squares (NLLS) is a widely-used method for modeling data. In NLLS, we wish to fit a model of the form,

$$\hat{y} = g(x, w)$$

where w is a vector of parameters and x is the vector of predictors. We find w by minimizing a least-squares function

$$f(w) = \sum_i (y_i - g(x_i, w))^2$$

where the summation is over training samples (x_i, y_i) . This is similar to linear least-squares, but the function $g(x, w)$ may not be linear in w . In general, this optimization has no closed-form expression. So numerical optimization must be used.

In this lab, we will implement gradient descent on NLLS in a problem of physical modeling of materials. Specifically, we will estimate parameters for expansion of copper as a function of temperature using a real dataset. In doing this lab, you will learn to: * Set up a nonlinear least squares as an unconstrained optimization function * Compute initial parameter estimates for a simple rational model * Compute the gradients of the least squares objective * Implement gradient descent for minimizing the objective * Implement momentum gradient descent * Visualize the convergence of the algorithm

We first import some key packages.

```
[ ]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import Ridge, LinearRegression
```

1.1 Load the Data

The NIST agency has an excellent [nonlinear regression website](#) that has several datasets for nonlinear regression problems. In this lab, we will use the data from a NIST study involving the thermal expansion of copper. The response variable is the coefficient of thermal expansion, and the predictor variable is temperature in degrees kelvin.

Hahn, T., NIST (1979), Copper Thermal Expansion Study. (unpublished)

You can download the data as follows.

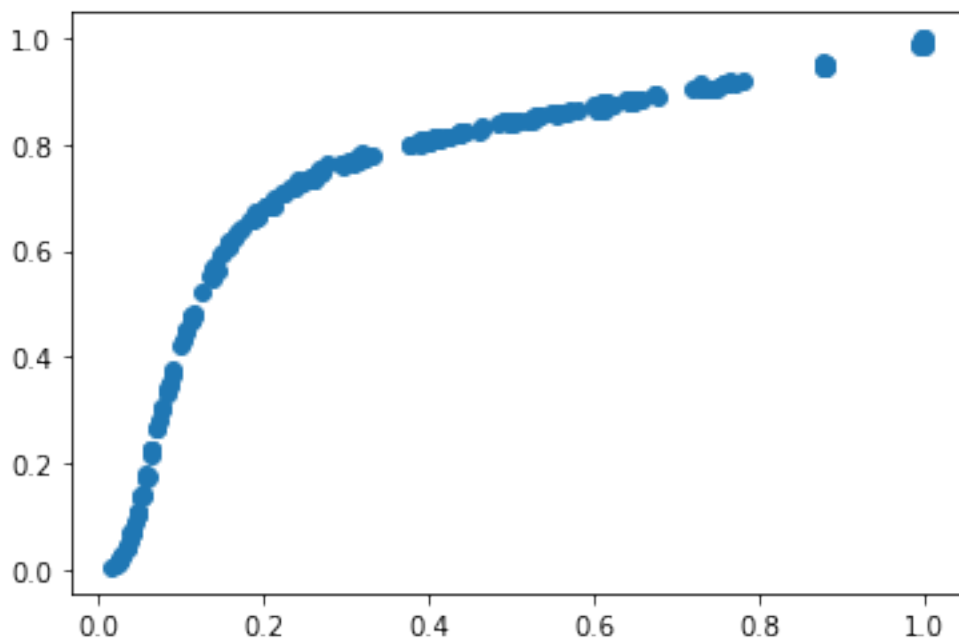
```
[ ]: url = 'https://itl.nist.gov/div898/strd/nls/data/LINKS/DATA/Hahn1.dat'
df = pd.read_csv(url, skiprows=60, sep=' ', skipinitialspace=True,
names=['y0', 'x0', 'dummy'])
df.head()
```

```
[ ]:      y0      x0  dummy
0  0.591  24.41   NaN
1  1.547  34.82   NaN
2  2.902  44.09   NaN
3  2.894  45.07   NaN
4  4.703  54.98   NaN
```

Extract the `x0` and `y0` into arrays. Rescale, `x0` and `y0` to values between 0 and 1 by dividing `x0` and `y0` by the maximum value. Store the scaled values in vectors `x` and `y`. The rescaling will help with the conditioning of the fitting. Plot, `y` vs. `x`.

```
[ ]: # TODO 1
x0 = np.array(df['x0'])
y0 = np.array(df['y0'])
x = x0/np.max(x0)
y = y0/np.max(y0)
plt.scatter(x, y)
```

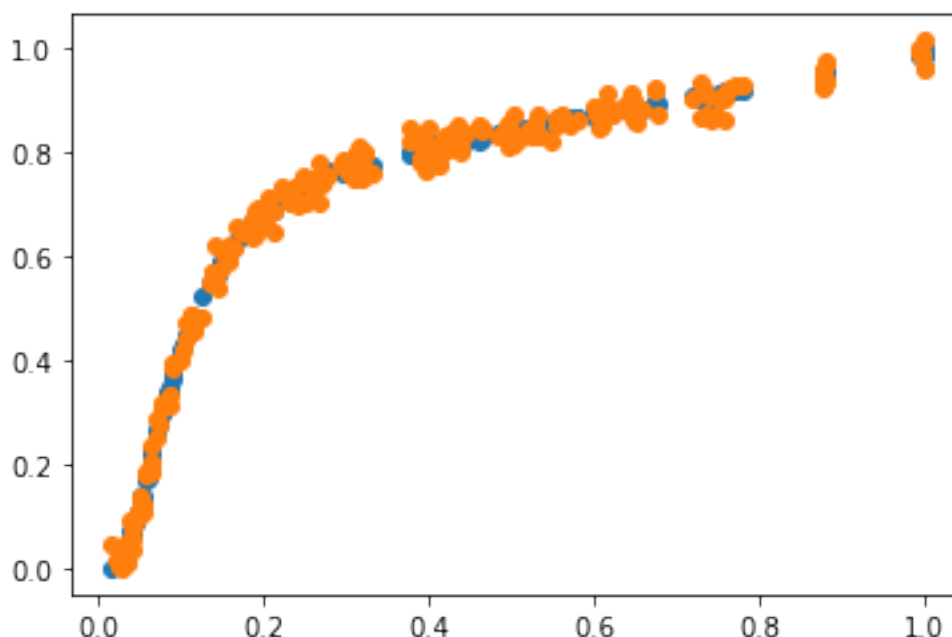
```
[ ]: <matplotlib.collections.PathCollection at 0x268b0d08fd0>
```



To make the problem a little more challenging, we will add some noise. Add random Gaussian noise with mean 0 and std. dev = 0.02 to y . Store the noisy results in yn . You can use the `np.random.normal()` function to add Gaussian noise. Plot yn vs. x .

```
[ ]: # TODO 2
yn = y + np.random.normal(0, 0.02, len(y))
plt.scatter(x, y)
plt.scatter(x, yn)
```

```
[ ]: <matplotlib.collections.PathCollection at 0x268b0ca9460>
```



Split the data (x, yn) into training and test. Let xtr, ytr be training data and xts, yts be the test data. You can use the `train_test_split` function. Set `test_size=0.33` so that 1/3 of the samples are held out for test.

```
[ ]: from sklearn.model_selection import train_test_split

# TODO 3
xtr, xts, ytr, yts = train_test_split(x, yn, test_size=0.33)
```

1.2 Initial Fit for a Rational Model

The [NIST website](#) suggests using a *rational* model of the form,

$$\hat{y} = (a[0] + a[1] * x + \dots + a[d] * x^d) / (1 + b[0] * x + \dots + b[d-1] * x^d)$$

with $d=3$. The model parameters are $\mathbf{w} = [a[0], \dots, a[d], b[0], \dots, b[d-1]]$ so there are $2d+1$ parameters total. Complete the function below that takes vectors \mathbf{w} and \mathbf{x} and predicts a set of values \mathbf{yhat} using the above model.

```
[ ]: def predict(w,x):

    # Get the length
    d = (len(w)-1)//2

    # TODO 4. Extract a and b from w
    a = w[:d + 1]
    b = w[d + 1:]
    b = np.concatenate([1], b))
    a = a[::-1]
    b = b[::-1]
    # TODO 5. Compute yhat. You may use the np.polyval function
    # But, remember you must flip the order the a and b
    Znum = np.polyval(a, x)
    Zden = np.polyval(b, x)
    yhat = Znum / Zden
    return yhat
```

When we fit with a nonlinear model, most methods only get convergence to a local minima. So, you need a good initial condition. For a rational model, one way to get is to realize that if:

$$y = (a[0] + a[1] * x + \dots + a[d] * x^d) / (1 + b[0] * x + \dots + b[d-1] * x^d)$$

Then:

$$y = a[0] + a[1] * x + \dots + a[d] * x^d - b[0] * x * y + \dots - b[d-1] * x^d * y$$

So, we can solve for the the parameters $\mathbf{w} = [\mathbf{a}, \mathbf{b}]$ from linear regression of the predictors,

$$Z[i, :] = [x[i], \dots, x[i]^d, y[i] * x[i], \dots, y[i] * x[i]^d]$$

```
[ ]: d = 3

# TODO 6. Create the transformed feature matrix
x_mat = xtr[:, None]
power = np.arange(1, d+1)
x_mat = x_mat**power
y_mat = -ytr[:, None]
y_mat = x_mat * y_mat
Z = np.concatenate((x_mat, y_mat), axis=1)

# TODO 7. Fit with parameters with linear regression
regr = LinearRegression()
```

```

regr.fit(Z, ytr)

# TODO 8
# Extract the parameters from regr.coef_ and regr.intercept_ and store the
# parameter vector in winit
winit = np.concatenate((regr.intercept_, regr.coef_))
print(winit)

```

```

[ -0.2415604   10.48120331 -38.057578    32.08864495   4.17713872
 -28.67493359  27.80691363]

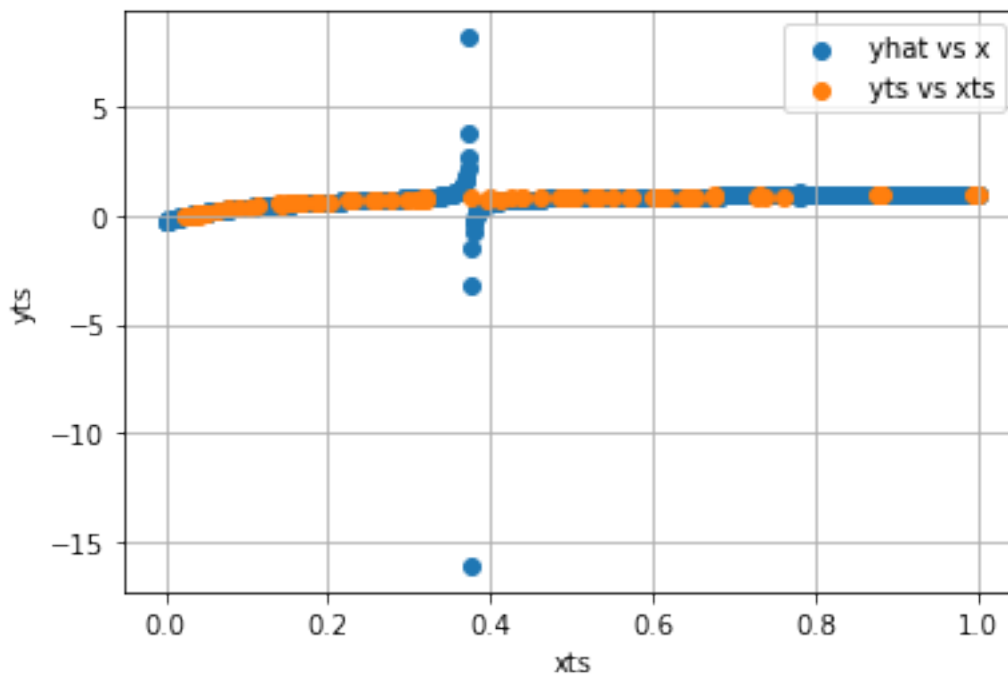
```

Now plot the predicted values of the \hat{y} vs. x using your estimated parameter `winit` for 1000 values x in $[0,1]$. On the same plot, plot y_{ts} vs. x_{ts} . You will see that you get a horrible fit.

```

[ ]: # TODO 9
xp = np.linspace(0, 1, 1000)
yhat = predict(winit, xp)
plt.scatter(xp, yhat)
# print(yhat)
plt.scatter(xts, yts)
plt.xlabel('xts')
plt.ylabel('yts')
plt.legend(['yhat vs x', 'yts vs xts'])
plt.grid()
plt.show()

```

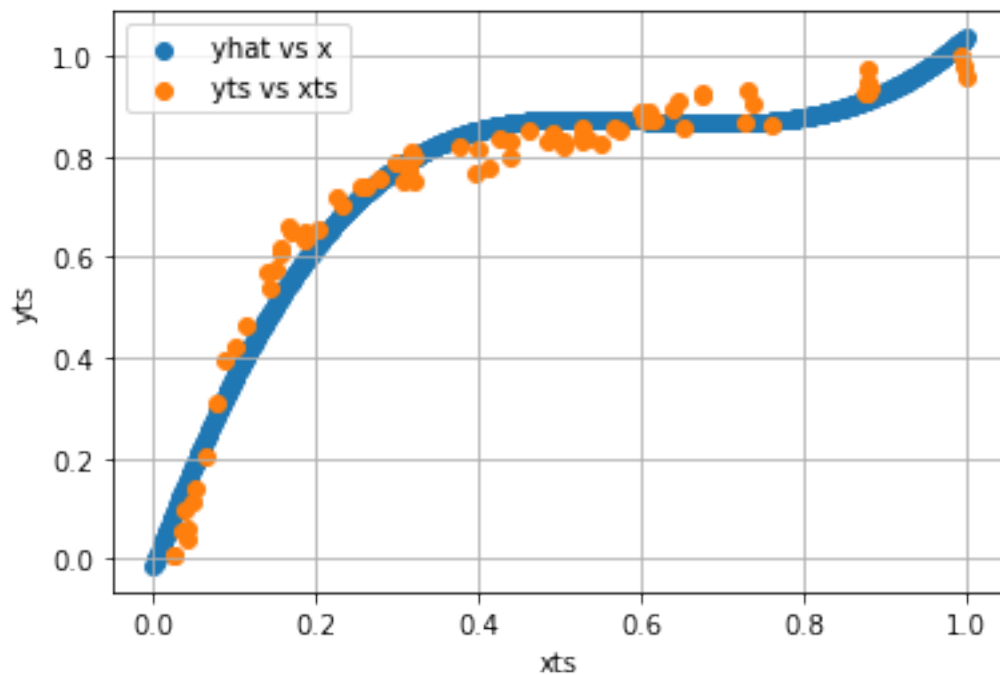


The reason the previous fit is poor is that the denominator in `yhat` goes close to zero. To avoid this problem, we can use Ridge regression, to try to keep the parameters close to zero. Re-run the fit above with Ridge with `alpha = 1e-3`. You should see you get a reasonable, but not perfect fit.

```
[ ]: # TODO 10. Fit with parameters with linear regression
regr = Ridge(alpha=1e-3)
regr.fit(Z, ytr)

# TODO 11
# Extract the parameters from regr.coef_ and regr.intercept_
winit = np.concatenate((regr.intercept_, regr.coef_))

# TODO 12
# Plot the results as above.
yhat = predict(winit, xp)
plt.scatter(xp, yhat)
plt.scatter(xts, yts)
plt.xlabel('xts')
plt.ylabel('yts')
plt.legend(['yhat vs x', 'yts vs xts'])
plt.grid()
plt.show()
```



1.3 Creating a Loss Function

We can now use gradient descent to improve our initial estimate. Complete the following function to compute

$$f(w) = 0.5 * \sum_i (y[i] - \hat{y}[i])^2$$

and `fgrad`, the gradient of `f(w)`.

```
[ ]: def feval(w,x,y):

    # TODO 13. Parse w
    d = (len(w)-1)//2
    a = w[:d + 1]
    b = w[d + 1:]
    b = np.concatenate([1], b))

    n = len(a)
    # TODO 14. Znum[i,j] = x[i]**j
    j = np.arange(n)
    Znum = x ** j[:, None]

    # TODO 15. Zden[i,j] = x[i]**(j+1)
    j = j + 1
    Zden = x ** j[:, None]

    # TODO 16. Compute yhat
    # Compute the numerator and denominator
    numer = Znum * a[:, None]
    deno = Zden * b[:, None]
    yhat = np.sum(numer) / np.sum(deno)

    # TODO 17. Compute loss
    yerr = y - yhat
    f = 0.5 * np.sum(yerr ** 2)

    # TODO 18. Compute gradients
    dy_da = Znum
    df_da = np.sum(-yerr * dy_da, axis=1)
    dy_db = Znum[1:] * y
    df_db = np.sum(yerr * dy_db, axis=1)
    # print(df_da.shape, df_db.shape)
    fgrad = np.hstack((df_da, df_db))

    return f, fgrad
```

Test the gradient function: * Take `w0=winit` and compute `f0,fgrad0 = feval(w0,xtr,ytr)` *

Take w_1 very close to w_0 and compute $f_1, f_{grad1} = \text{feval}(w_1, x_{tr}, y_{tr})$ * Verify that $f_1 - f_0$ is close to the predicted value based on the gradient.

```
[ ]: # TODO 19
w0 = winit
f0, fgrad0 = feval(w0, xtr, ytr)
w1 = winit + 1e-6
f1, fgrad1 = feval(w1, xtr, ytr)
print([f1 - f0, fgrad0 * (w1 - w0)])
```

```
[0.00040283932365525743, array([ 1.76370351e-04,  5.37493742e-05,
2.78300266e-05,  1.79820881e-05,
-4.30583962e-05, -2.43025810e-05, -1.62472801e-05])]
```

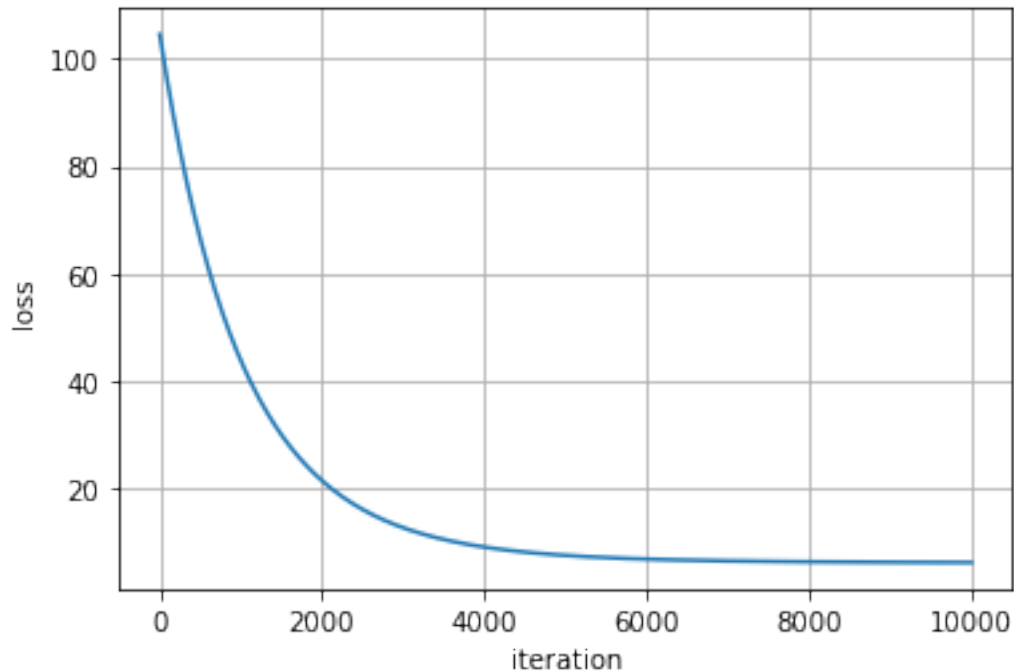
1.4 Implement gradient descent

We will now try to minimize the loss function with gradient descent. Using the function `feval` defined above, implement gradient descent. Run gradient descent with a step size of `alpha=1e-6` starting at `w=winit`. Run it for `nit=10000` iterations. Compute `fgd[it]`= the objective function on iteration `it`. Plot `fgd[it]` vs. `it`.

You should see that the training loss decreases, but it still hasn't converged after 10000 iterations.

```
[ ]: # TODO 20
# fgd = ...
nit = 10000
step = 1e-6
w = winit
fgd = np.zeros(nit)
for i in range(nit):
    f, fgrad = feval(w, xtr, ytr)
    w = w - step * fgrad
    fgd[i] = f

interval = np.linspace(1, nit, nit)
plt.plot(interval, fgd)
plt.xlabel('iteration')
plt.ylabel('loss')
plt.grid()
plt.show()
```

Now, try to get a faster convergence with adaptive step-size using the Armijo rule. Implement the gradient descent with adaptive step size. Let `fadapt[it]` be the loss function on iteration `it`. Plot `fadapt[it]` and `fgd[it]` vs. `it` on the same graph. You should see a slight improvement, but not much.

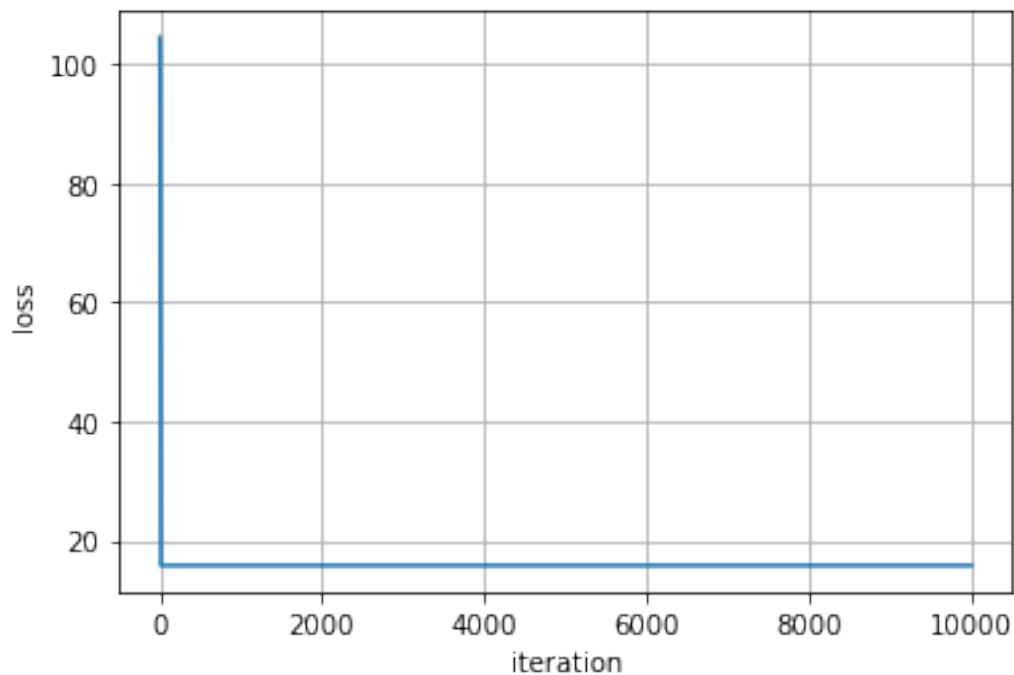
```
[ ]: # TODO 21
# fadapt = ...
nit = 10000
step = 1e-6 # Initial step

wAdap = winit
c = 1/2
fgd = np.zeros(nit)
f0, fgrad0 = feval(wAdap, xtr, ytr)
for it in range(nit):
    w1 = wAdap - step * fgrad0
    f1, fgrad1 = feval(w1, xtr, ytr)
    if ((f1-f0) < (c * fgrad0.dot(w1-w0))) and (f1 < f0):
        step = step / c
        f0 = f1
        fgrad0 = fgrad1
        wAdap = w1
    else:
        step = step * c
    fgd[it] = f0
```

```

interval = np.linspace(1, nit, nit)
plt.plot(interval, fgd)
plt.xlabel('iteration')
plt.ylabel('loss')
plt.grid()
plt.show()

```

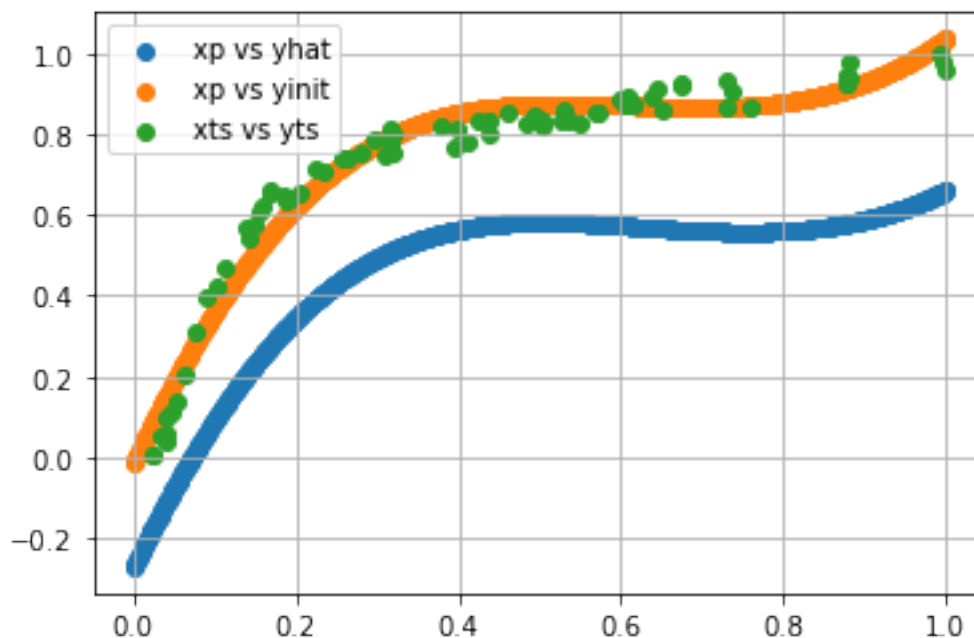


Using the final estimate for w from the adaptive step-size plot the predicted values of the \hat{y} vs. x for 1000 values x in $[0,1]$. On the same plot, plot \hat{y} vs. x for the initial parameter $w=w_{init}$. Also, plot y_{ts} vs. x_{ts} . You should see that gradient descent was able to improve the estimate slightly, although the initial estimate was not too bad.

```

[ ]: # TODO 22
xp = np.linspace(0, 1, 1000)
yhat = predict(wAdap, xp)
yinit = predict(winit, xp)
plt.scatter(xp, yhat)
plt.scatter(xp, yinit)
plt.scatter(xts, yts)
plt.legend(['xp vs yhat', 'xp vs yinit', 'xts vs yts'])
plt.grid()
plt.show()

```



1.5 Momentum Gradient Descent

This section is bonus.

One way to improve gradient descent is to use *momentum*. In momentum gradient descent, the update rule is:

```
f, fgrad = feval(w,...)
z = beta*z + fgrad
w = w - step*z
```

This is similar to gradient descent, except that there is a second order term on the gradient. Implement this algorithm with `beta = 0.99` and `step=1e-3`. Compare the convergence of the loss function with gradient descent.

```
[ ]: # TODO 23
nit = 1000
step = 1e-3
beta = 0.99
```

```
[ ]: # TODO 24
# plot yhat vs. x
```

1.6 Beyond This Lab

In this lab, we have just touched at some of the ideas in optimization. There are several other important algorithms that you can explore: * [Levenberg-Marquardt](#) method for non-linear least squares * Newton's method * More difficult non-linear least squares problems.

[]: