6 Fitting Models

Problem 6.1 Show that the derivates of the least squares loss function in equation 6.5 are given by the expressions in equation 6.7.

The least squares loss function $L[\phi]$ for 1D regression model is given by

$$L[\phi] = \sum_i (\phi_0 + \phi_1 x_i - y_i)^2$$

Partial derivative w.r.t. ϕ_0 is:

$$\begin{aligned}\frac{\partial L[\phi]}{\partial \phi_0} &= \sum_i 2(\phi_0 + \phi_1 x_i - y_i)\frac{\partial}{\partial \phi_0}(\phi_0 + \phi_1 x_i - y_i) \\ &= \sum_i 2(\phi_0 + \phi_1 x_i - y_i)\end{aligned}$$

Partial derivate w.r.t. ϕ_1 is:

$$\begin{aligned}\frac{\partial L[\phi]}{\partial \phi_1} &= \sum_i 2(\phi_0 + \phi_1 x_i - y_i)\frac{\partial}{\partial \phi_1}(\phi_0 + \phi_1 x_i - y_i) \\ &= \sum_i 2(\phi_0 + \phi_1 x_i - y_i)x_i\end{aligned}$$

Problem 6.2 A surface is convex if the eigenvalues of the Hessian $\mathbf{H}[\boldsymbol{\phi}]$ are positive everywhere. In this case, the surface has a unique minimum, and optimization is easy. Find an algebraic expression for the Hessian matrix,

$$\mathbf{H}[\boldsymbol{\phi}] = \begin{bmatrix} \frac{\partial^2 L}{\partial \phi_0^2} & \frac{\partial^2 L}{\partial \phi_0 \partial \phi_1} \\ \frac{\partial^2 L}{\partial \phi_1 \partial \phi_0} & \frac{\partial^2 L}{\partial \phi_1^2} \end{bmatrix}$$

for the linear regression model (equation 6.5). Prove that this function is convex by showing that the eigenvalues are always positive. This can be done by showing that both the trace and the determinant of the matrix are positive.

Each elements of the Hessian matrix are computed as:

$$
\begin{aligned}
\frac{\partial^2 L}{\partial \phi_0^2} &= \sum_i 2 = 2I \\
\frac{\partial^2 L}{\partial \phi_0 \partial \phi_1} &= \sum_i 2x_i \\
\frac{\partial^2 L}{\partial \phi_1 \partial \phi_0} &= \sum_i 2x_i \\
\frac{\partial^2 L}{\partial \phi_1^2} &= \sum_i 2x_i^2
\end{aligned}
$$

The trace of the hessian matrix is thus

$$\text{trace}\, \mathbf{H}[\boldsymbol{\phi}] = 2I + 2\sum_i x_i^2 > 0$$

and the determinant is also non-negative.

$$\det \mathbf{H}[\boldsymbol{\phi}] = (2I)(2\sum_i x_i^2) - 4(\sum_i x_i)^2 \geq 0$$

This inequality holds by the Cauchy-Schwarz inequality:

$$(1^2 + \cdots + 1^2)(x_1^2 + \cdots + x_I^2) \geq (x_1 + \cdots + x_I)^2$$

Problem 6.3 Compute the derivatives of the least squares loss $L[\boldsymbol{\phi}]$ with respect to the parameters ϕ_0 and ϕ_1 for the Gabor model (equation 6.8)

The Gabor model is defined by

$$f[x, \boldsymbol{\phi}] = \sin[\phi_0 + 0.06 \cdot \phi_1 x] \cdot \exp\left(-\frac{(\phi_0 + 0.06 \cdot \phi_1 x)^2}{32.0}\right)$$

Note that

$$
\begin{aligned}
\frac{\partial}{\partial \phi_0} f[x, \boldsymbol{\phi}] = \frac{\partial}{\partial \phi_0} & \sin[\phi_0 + 0.06 \cdot \phi_1 x] \cdot \exp\left(-\frac{(\phi_0 + 0.06 \cdot \phi_1 x)^2}{32.0}\right) \\
&= \cos[\phi_0 + 0.06 \cdot \phi_1 x] \cdot \exp\left(-\frac{(\phi_0 + 0.06 \cdot \phi_1 x)^2}{32.0}\right) \\
&+ \sin[\phi_0 + 0.06 \cdot \phi_1 x] \cdot \exp\left(-\frac{(\phi_0 + 0.06 \cdot \phi_1 x)^2}{32.0}\right) \cdot \left(-\frac{\phi_0 + 0.06 \cdot \phi_1 x}{16.0}\right)
\end{aligned}
$$

and

$$\begin{aligned}\frac{\partial}{\partial \phi_1} f[x, \boldsymbol{\phi}] &= \frac{\partial}{\partial \phi_1} \sin[\phi_0 + 0.06 \cdot \phi_1 x] \cdot \exp\left(-\frac{(\phi_0 + 0.06 \cdot \phi_1 x)^2}{32.0}\right) \\ &= \cos[\phi_0 + 0.06 \cdot \phi_1 x] \cdot (0.06x) \cdot \exp\left(-\frac{(\phi_0 + 0.06 \cdot \phi_1 x)^2}{32.0}\right) \\ &+ \sin[\phi_0 + 0.06 \cdot \phi_1 x] \cdot \exp\left(-\frac{(\phi_0 + 0.06 \cdot \phi_1 x)^2}{32.0}\right) \cdot \left(-\frac{\phi_0 + 0.06 \cdot \phi_1 x}{16.0}\right) \cdot (0.06x)\end{aligned}$$

Partial derivatives of the least squares loss are:

$$\frac{\partial L}{\partial \phi_0} = 2 \sum_i (f[x_i, \boldsymbol{\phi}] - y_i) \frac{\partial}{\partial \phi_0} f[x_i, \boldsymbol{\phi}]$$
$$\frac{\partial L}{\partial \phi_1} = 2 \sum_i (f[x_i, \boldsymbol{\phi}] - y_i) \frac{\partial}{\partial \phi_1} f[x_i, \boldsymbol{\phi}]$$

Problem 6.4 The logistic regression model use a linear function to assign an input $\mathbf{x}$ to one of two classes $y \in \{0, 1\}$. For a 1D input and a 1D output, it has two parameters, ϕ_0 and ϕ_1, and is defined by:

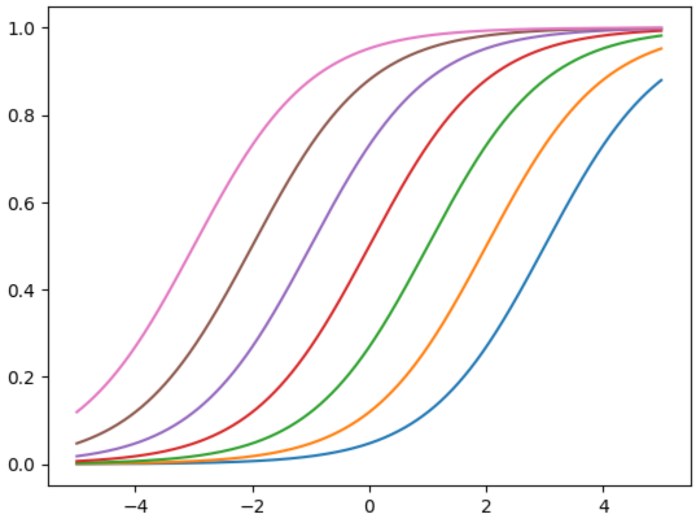
$$Pr(y = 1|x) = \text{sig}[\phi_0 + \phi_1 x]$$

where $\text{sig}[\cdot]$ is the logistic sigmoid function:

$$\text{sig}[z] = \frac{1}{1 + \exp[-z]}$$

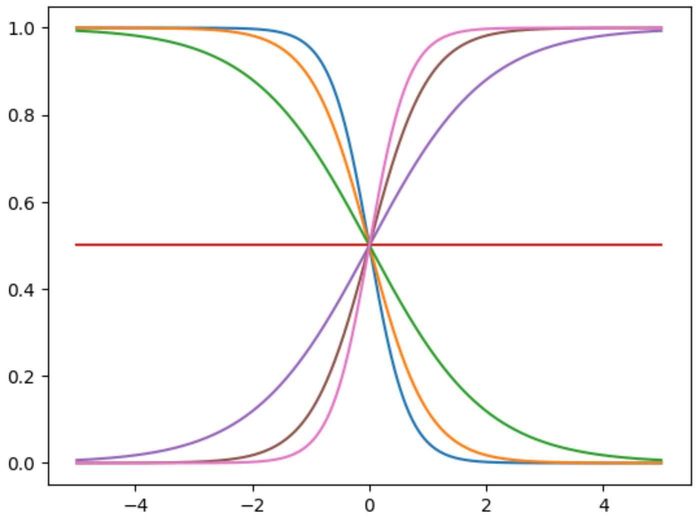
(i) Plot y against x for this model for different values of ϕ_0 and ϕ_1 and explain the qualitative meaning of each parameter.

Following is the plot with different ϕ_0 values with fixed $\phi_1 = 1$:



ϕ_0 controls the shift of the sigmoid function.

And the plot with different ϕ_1 values with fixed $\phi_0 = 0$:



ϕ_1 controls the steepness of the sigmoid path.

(ii) What is a suitable loss function for this model?

We would attempt to minimize the negative log-likelihood given the logistic regression model.

$$
\begin{aligned}
L[\boldsymbol{\phi}] &= -\sum_i y_i \log Pr(y=1|x_i) + (1-y_i)\log(1-Pr(y=1|x_i))) \\
&= -\sum_i y_i \log \mathrm{sig}[\phi_0+\phi_1 x] + (1-y_i)\log(1-\mathrm{sig}[\phi_0+\phi_1 x])
\end{aligned}
$$

(iii) Compute the derivates of this loss function with respect to the parameters.

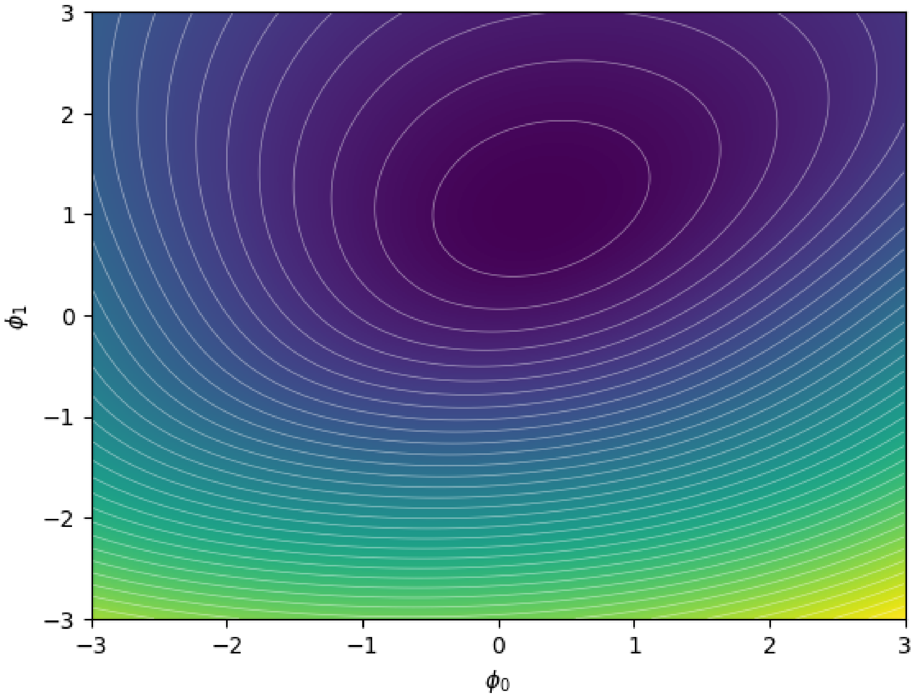
The derivate of the sigmoid function is:

$$
\frac{\partial}{\partial z}\mathrm{sig}[z] = \frac{\exp[-z]}{(1+\exp[-z])^2}
$$

Thus, the partial derivatives of the loss is

$$
\begin{aligned}
\frac{\partial L}{\partial \phi_0} &= -\sum_i \left(\frac{y_i}{\mathrm{sig}[\phi_0+\phi_1 x_i]} - \frac{1-y_i}{1-\mathrm{sig}[\phi_0+\phi_1 x_i]}\right)\frac{\exp[-(\phi_0+\phi_1 x_i)]}{(1+\exp[-(\phi_0+\phi_1 x_i)])^2} \\
\frac{\partial L}{\partial \phi_1} &= -\sum_i \left(\frac{y_i}{\mathrm{sig}[\phi_0+\phi_1 x_i]} - \frac{1-y_i}{1-\mathrm{sig}[\phi_0+\phi_1 x_i]}\right)\frac{x_i\exp[-(\phi_0+\phi_1 x_i)]}{(1+\exp[-(\phi_0+\phi_1 x_i)])^2}
\end{aligned}
$$

(iv) Generate ten data points from a normal distribution with mean -1 and standard deviation 1 and assign them the label $y=0$. Generate another ten data points from a normal distribution with mean 1 and standard deviation 1 and assign these the label $y=1$. Plot the loss as a heatmap in terms of the two parameters ϕ_0 and ϕ_1.



(v) Is this loss function convex? How could you prove this?

I would not rather compute Hessian matrix by myself…

But here is a definite proof. https://math.stackexchange.com/a/3198681

First, we prove that the functions $-\log \text{sig}[z]$ and $-\log(1-\text{sig}[z])$ are convex.

This can be done by computing the derivatives:

$$\frac{\partial}{\partial z}(-\log \text{sig}[z]) = \frac{\partial}{\partial z}\log(1+\exp[-z]) = \frac{-\exp[-z]}{1+\exp[-z]} = \text{sig}[z]-1$$

Since the derivative is increasing, the second-order derivative is everywhere positive, thus $-\log \text{sig}[z]$ is convex.

Similarly,

$$\frac{\partial}{\partial z}(-\log(1-\text{sig}[z])) = \frac{\partial}{\partial z}(\log(1+\exp[-z])+z) = \text{sig}[z]$$

likewise $-\log(1-\text{sig}[z])$ is also convex.

Next, we prove that if f is convex, then $f(ax+b)$ is also convex. The proof is simple since $f'' = a^2 f'' > 0$.

Notice that the loss for the logistic regression is the linear combination of $-\log \text{sig}[\phi_0 + \phi_1 x]$ and $-\log(1 - \text{sig}[\phi_0 + \phi_1 x])$. By the previous claims, the overall loss function should be also convex w.r.t. ϕ_0, ϕ_1.

Problem 6.5 Compute the derivatives of the least squares loss with respect to the ten parameters of the simple neural network model introduced in equation 3.1:

$$f[x, \boldsymbol{\phi}] = \phi_0 + \phi_1 a[\theta_{10} + \theta_{11}x] + \phi_2 a[\theta_{20} + \theta_{21}x] + \phi_3 a[\theta_{30} + \theta_{31}x]$$

Think carefully about what the derivative of the ReLU function $a[\cdot]$ will be.

The derivative of the ReLU function is 1 for $x > 0$ and 0 elsewhere, which is a heaviside step function $h[\cdot]$. Thus we can compute the partial derivatives for the ten parameters as follows:

$$\begin{aligned}
\frac{\partial f}{\partial \phi_0} &= 1 \\
\frac{\partial f}{\partial \phi_1} &= a[\theta_{10} + \theta_{11}x] \\
\frac{\partial f}{\partial \phi_2} &= a[\theta_{20} + \theta_{21}x] \\
\frac{\partial f}{\partial \phi_3} &= a[\theta_{30} + \theta_{31}x] \\
\frac{\partial f}{\partial \theta_{10}} &= \phi_1 \cdot h[\theta_{10} + \theta_{11}x] \\
\frac{\partial f}{\partial \theta_{11}} &= \phi_1 x \cdot h[\theta_{10} + \theta_{11}x] \\
\frac{\partial f}{\partial \theta_{20}} &= \phi_1 \cdot h[\theta_{20} + \theta_{21}x] \\
\frac{\partial f}{\partial \theta_{21}} &= \phi_1 x \cdot h[\theta_{20} + \theta_{21}x] \\
\frac{\partial f}{\partial \theta_{30}} &= \phi_1 \cdot h[\theta_{30} + \theta_{31}x] \\
\frac{\partial f}{\partial \theta_{31}} &= \phi_1 x \cdot h[\theta_{30} + \theta_{31}x]
\end{aligned}$$

Problem 6.6 Which of the functions in figure 6.11 is convex? Justify your answer. Characterize each of the points 1-7 as (i) a local minimum, (ii) the global minimum, or (iii) neither.

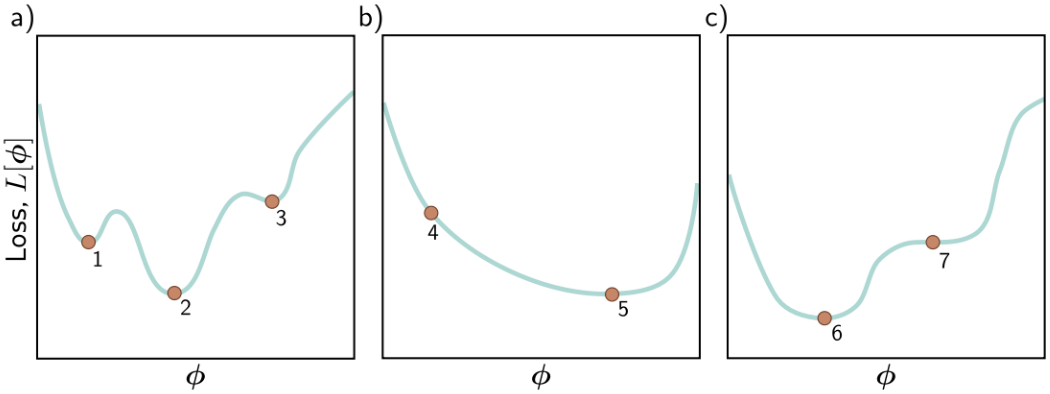


Figure 6.11 Three 1D loss functions for problem 6.6.

Only (b) is a convex function.

(i) Local minimum: points 1, 2, 3, 5, and 6.

(ii) Global minimum: points 2, 5, and 6.

(iii) Neither: points 4 and 7.

Problem 6.7 The gradient descent trajectory for path 1 in figure 6.5a oscillates back and forth inefficiently as it moves down the valley toward the minimum. It's also notable that it turns at right angles to the previous direction at each step. Provide a qualitative explanation for these phenomena. Propose a solution that might help prevent this behavior.

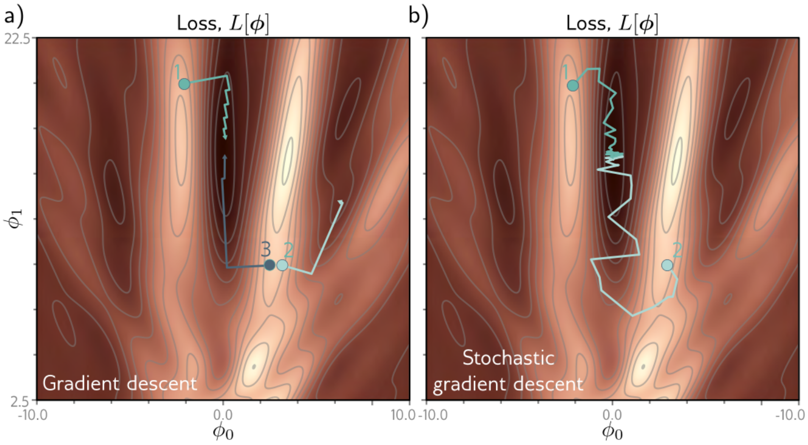


Figure 6.5 Gradient descent vs. stochastic gradient descent. a) Gradient descent with line search. As long as the gradient descent algorithm is initialized in the right "valley" of the loss function (e.g., points 1 and 3), the parameter estimate will move steadily toward the global minimum. However, if it is initialized outside this valley (e.g., point 2), it will descend toward one of the local minima. b) Stochastic gradient descent adds noise to the optimization process, so it is possible to escape from the wrong valley (e.g., point 2) and still reach the global minimum.

When the gradient descent algorithm arrives at some point, it finds the minimum along the gradient of the current position. At such position, the next trajectory should be at right angle: if not, the position would not be the minimal point along the direction. This causes the overall trajectory to oscillate and be inefficient. We can introduce the concept of momentum to average out those oscillations.

Problem 6.8 Can (non-stochastic) gradient descent with a *fixed* learning rate escape local minima?

If the movement in the position is large enough to overshoot the local valley, then the algorithm could escape the local minima.

Problem 6.9 We run the stochastic gradient descent algorithm for 1,000 iterations on a dataset of size 100 with a batch size of 20. For how many epochs did we train the model?

5 iterations for iterating through the entire dataset, so 200 epochs.

Problem 6.10 Show that the momentum term $\mathbf{m}_t$ (equation 6.11) is an infinite weight sum of the gradients at the previous iterations and derive an expression for the coefficients (weights) of that sum.

The recursive definition of $\mathbf{m}_t$ is

$$\mathbf{m}_{t+1} \leftarrow \beta \cdot \mathbf{m}_t + (1-\beta)\mathbf{l}_t$$

where $\mathbf{l}_t$ is the gradient at t-th iteration $\sum_{i \in \mathcal{B}_t} \frac{\partial l_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$.

We can expand this expression to show the full form:

$$\begin{aligned}\mathbf{m}_t =& (1-\beta) \cdot \mathbf{l}_{t-1} \\ &+ (1-\beta)\beta \cdot \mathbf{l}_{t-2} \\ &+ (1-\beta)\beta^2 \cdot \mathbf{l}_{t-3} \\ &+ \cdots\end{aligned}$$

Problem 6.11 What dimensions will be the Hessian have if the model has one million parameters?

10^{12} parameters.