1. Assume that we have some data $x_1, \ldots x_n \in \mathbb{R}$. Our goal is to find a constant b such that $\sum_i (y_i - b)^2$ is minimized.

    1. Find an analytic solution for the optimal value of b.
    2. How does this problem and its solution relate to the normal distribution?
    3. What if we change the loss from

    $$\sum_i (x_i - b)^2$$

    to

    $$\sum_i |x_i - b|?$$

    Can you find the optimal solution for b?

2. Prove that the affine functions that can be expressed by $y = Wx + b$ are equivalent to linear functions on $(x, 1)$.

3. Assume that you want to find quadratic functions of y, i.e.,

$$y = Wx^2 + b.$$

How would you formulate this in a deep network?

4. Recall that one of the conditions for the linear regression problem to be solvable was that the design matrix X has full rank.

    1. What happens if this is not the case?
    2. How could you fix it? What happens if you add a small amount of coordinate-wise independent Gaussian noise to all entries of X?
    3. What is the expected value of the design matrix $X^t X$ in this case?
    4. What happens with stochastic gradient descent when $X^t X$ does not have full rank?

5. Assume that the noise model governing the additive noise e_i is the exponential distribution. That is, $p(e_i) = \lambda e^{-\lambda e_i}$.

    1. Write out the negative log-likelihood of the data under the model $p(e_i)$.
    2. Can you find a closed form solution?
    3. Suggest a minibatch stochastic gradient descent algorithm to solve this problem. What could possibly go wrong (hint: what happens near the stationary point as we keep on updating the parameters)? Can you fix this?

6. Assume that we want to design a neural network with two layers by composing two linear layers. That is, the output of the first layer becomes the input of the second layer. Why would such a naive composition not work?

7. What happens if you want to use regression for realistic price estimation of houses or stock prices?

    1. Show that the additive Gaussian noise assumption is not appropriate. Hint: can we have negative prices? What about fluctuations?
    2. Why would regression to the logarithm of the price be much better, i.e., $y = \log \text{price}$?
    3. What do you need to worry about when dealing with pennystock, i.e., stock with very low prices? Hint: can you trade at all possible prices? Why is this a bigger problem for cheap stock?
    4. For more information review the celebrated Black-Scholes model for option pricing (Black and Scholes, 1973).

8. Suppose we want to use regression to estimate the number of apples sold in a grocery store.

    1. What are the problems with a Gaussian additive noise model? Hint: you are selling apples, not oil.
    2. The Poisson distribution captures distributions over counts. It is given by $P(k; \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$. Here λ is the rate function and k is the number of events you see. Prove that λ is the expected value of counts k.
    3. Design a loss function associated with the Poisson distribution.
    4. Design a loss function for estimating $\log \lambda$ instead.