9 Regularization

Problem 9.1 Consider a model where the prior distribution over the parameters is a normal distribution with mean zero and variance $\sigma^2_{\boldsymbol{\phi}}$ so that

$$Pr(\boldsymbol{\phi}) = \prod_{j=1}^{J} \text{Norm}_{\phi_j}[0, \sigma^2_{\boldsymbol{\phi}}]$$

where j indexes the model parameters. We now maximize $\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\phi})Pr(\boldsymbol{\phi})$. Show that the associated loss function of this model is equivalent to L2 regularization.

Consider the negative log-likelihood of the objective function.

$$L[\boldsymbol{\phi}] = -\sum_{i=1}^{I} \log Pr(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\phi}) - \log Pr(\boldsymbol{\phi})$$

The first term of the loss function corresponds to the loss term constructed from maximum likelihoods criterion without knowledge of the prior distribution over the parameters. The second term $-\log Pr(\boldsymbol{\phi})$ is the regularization term

$$\begin{aligned}
-\log Pr(\boldsymbol{\phi}) &= -\sum_{j=1}^{J} \log[\text{Norm}_{\phi_j}[0, \sigma^2_{\boldsymbol{\phi}}]] \\
&= \sum_{j=1}^{J} [\frac{1}{2\sigma^2_{\boldsymbol{\phi}}} \phi_j^2 + \log \sqrt{2\pi}\sigma_{\boldsymbol{\phi}}] \\
&= \lambda \cdot \sum_{j=1}^{J} \phi_j^2 + C
\end{aligned}$$

which is equivalent to the L2 regularization.

Problem 9.2 How do the gradients of the loss function change when L2 regularization (equation 9.5) is added?

The gradient of the L2 regularization term contributes the model weights to converge towards the origin. Thus the gradient of the loss function is adjusted with the additional term heading towards the origin.

Problem 9.3 Consider a linear regression model $y = \phi_0 + \phi_1 x$ with input x, output y, and parameters ϕ_0 and ϕ_1. Assume we have I training examples $\{x_i, y_i\}$ and use a least squares loss. Consider adding Gaussian noise with mean zero and variance σ_x^2 to the inputs x_i at each training iteration. What is the expected gradient update?

The gradient update for parameters ϕ_0, ϕ_1 is as follows:

$$\frac{\partial L}{\partial \phi_0} = \sum_i 2(\phi_0 + \phi_1 x_i' - y_i)$$
$$\frac{\partial L}{\partial \phi_1} = \sum_i 2x_i'(\phi_0 + \phi_1 x_i' - y_i)$$

where x_i' indicate the inputs with the noise.

For the partial gradient w.r.t. ϕ_0, we can expect that the noises will cancel out and be identical to the gradient without the noise. However, partial gradient w.r.t. ϕ_1 contains a quadratic term $2\phi_1 x_i'^2$, which will result to an additional term $2\phi_1\sigma_x^2$ to the gradient update for each training example.

Problem 9.4 Derive the loss function for multiclass classification when we use label smoothing so that the target probability distribution has 0.9 at the correct class and the remaining probability mass of 0.1 is divided between the remaining $D_o - 1$ classes.

We choose the categorical distribution with label smoothing applied

$$Pr(y_i|\mathbf{f}[\mathbf{x}_i, \boldsymbol{\phi}]) = 0.9 \cdot \text{softmax}_{y_i}[\mathbf{f}_{y_i}[\mathbf{x}_i, \boldsymbol{\phi}]] + \frac{0.1}{D_o - 1}\sum_{k \neq y_i} \text{softmax}_k[\mathbf{f}_k[\mathbf{x}_i, \boldsymbol{\phi}]]$$

so that the probability for label y_i is 0.9 and $0.1/(D_o - 1)$ for all the other labels.

Therefore, the loss function derived from the negative log likelihood function would be:

$$L[\boldsymbol{\phi}] = -\sum_i \left[\log[0.9 \exp_{y_i}[\mathbf{f}_{y_i}[\mathbf{x}_i, \boldsymbol{\phi}]] + \frac{0.1}{D_o - 1}\sum_{k \neq y_i} \exp_k[\mathbf{f}_k[\mathbf{x}_i, \boldsymbol{\phi}]]] - \log[\sum_k \exp_k[\mathbf{f}_k[\mathbf{x}_i, \boldsymbol{\phi}]]]\right]$$

Problem 9.5 Show that the weight decay parameter update with decay rate λ:

$$\boldsymbol{\phi} \leftarrow (1 - \lambda)\boldsymbol{\phi} - \alpha\frac{\partial L}{\partial \boldsymbol{\phi}}$$

on the original loss function $L[\boldsymbol{\phi}]$ is equivalent to a standard gradient update using L2 regularization so that the modified loss function $\tilde{L}[\boldsymbol{\phi}]$ is:

$$\tilde{L}[\boldsymbol{\phi}] = L[\boldsymbol{\phi}] + \frac{\lambda}{2\alpha}\sum_k \phi_k^2$$

where $\boldsymbol{\phi}$ are the parameters, and α is the learning rate.

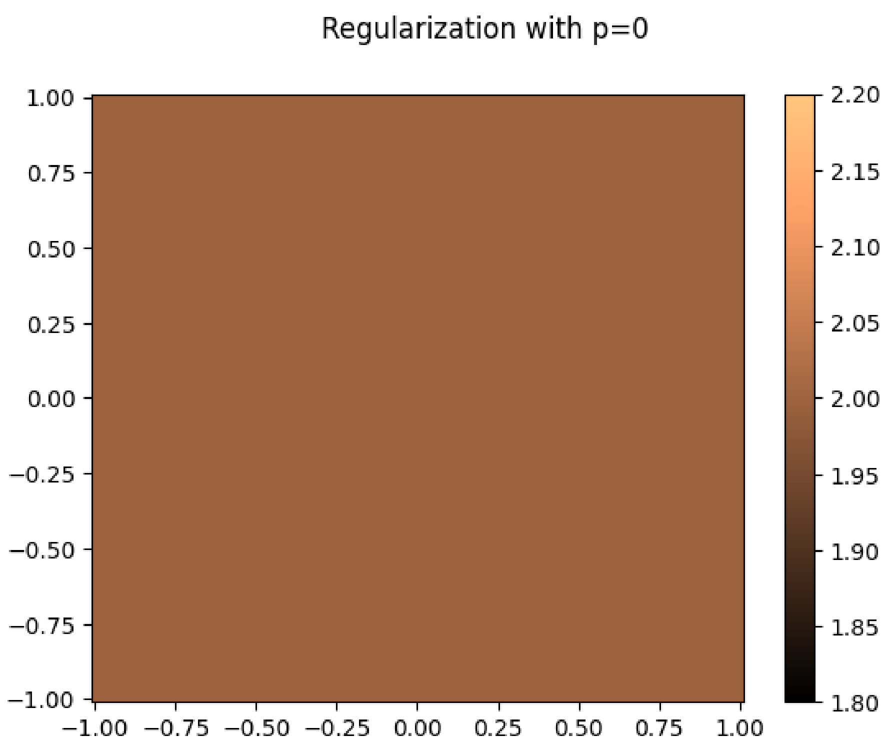
We can see that the gradient of the modified loss function is

$$\frac{\partial \tilde{L}}{\partial \boldsymbol{\phi}} = \frac{\partial L}{\partial \boldsymbol{\phi}} + \frac{\lambda}{\alpha}\boldsymbol{\phi}$$

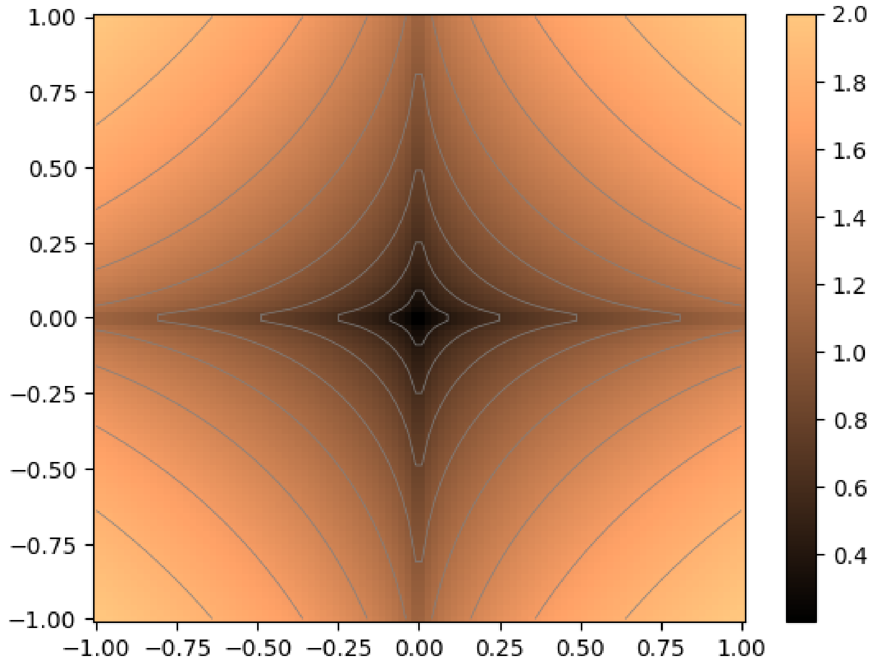
so the GD update step of this loss would be

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \alpha\frac{\partial \tilde{L}}{\partial \boldsymbol{\phi}} = (1-\lambda)\boldsymbol{\phi} - \alpha\frac{\partial L}{\partial \boldsymbol{\phi}}$$

Problem 9.6 Consider a model with parameters $\boldsymbol{\phi} = [\phi_0, \phi_1]^T$. Draw the L0, $L^{1/2}$, and L1 regularization terms in a similar form to figure 9.1b. The LP regularization term is $\sum_d |\phi_d|^P$.



Regularization with p=0.5



Regularization with p=1