Neural network regression and ALE plot

Components of neural network regression

- Architecture: how many layers, what kind of activation functions, etc.
- Parameters: architecture and parameters together characterize a neural network;
- Loss function: describes how well the model that is trained fits the data, to obtain a well-fit model, we want to minimize this loss function;
- Optimizer: you need to choose the optimization method either when using python function MLPRegressor or construct by yourself;

Training a neural network regressor

1. Once defined the architecture, parameters will be first assigned with random values;

2. During the training, take stochastic gradient descent for example, a random input $\boldsymbol{X}_i$ is fed into the model, the model generates the corresponding output $\hat{y}_i$ in the forward direction;

3. The loss function evaluates how far $\hat{y}_i$ is from y_i;

4. The optimization method evaluates the derivative of parameters w.r.t. the loss function in a backward manner, and updates each parameter;

5. Iterate the 2-4 procedure over the total dataset, and this is one epoch; run enough number of epochs, till some criterion is achieved.

Global vs. local optimization

Suppose we want to minimize an arbitrary function, $g(\boldsymbol{w})$,

- Global optimization: evaluate the function using a large number of input points, and treat the one with minimum function value as the global minimizer of the function;

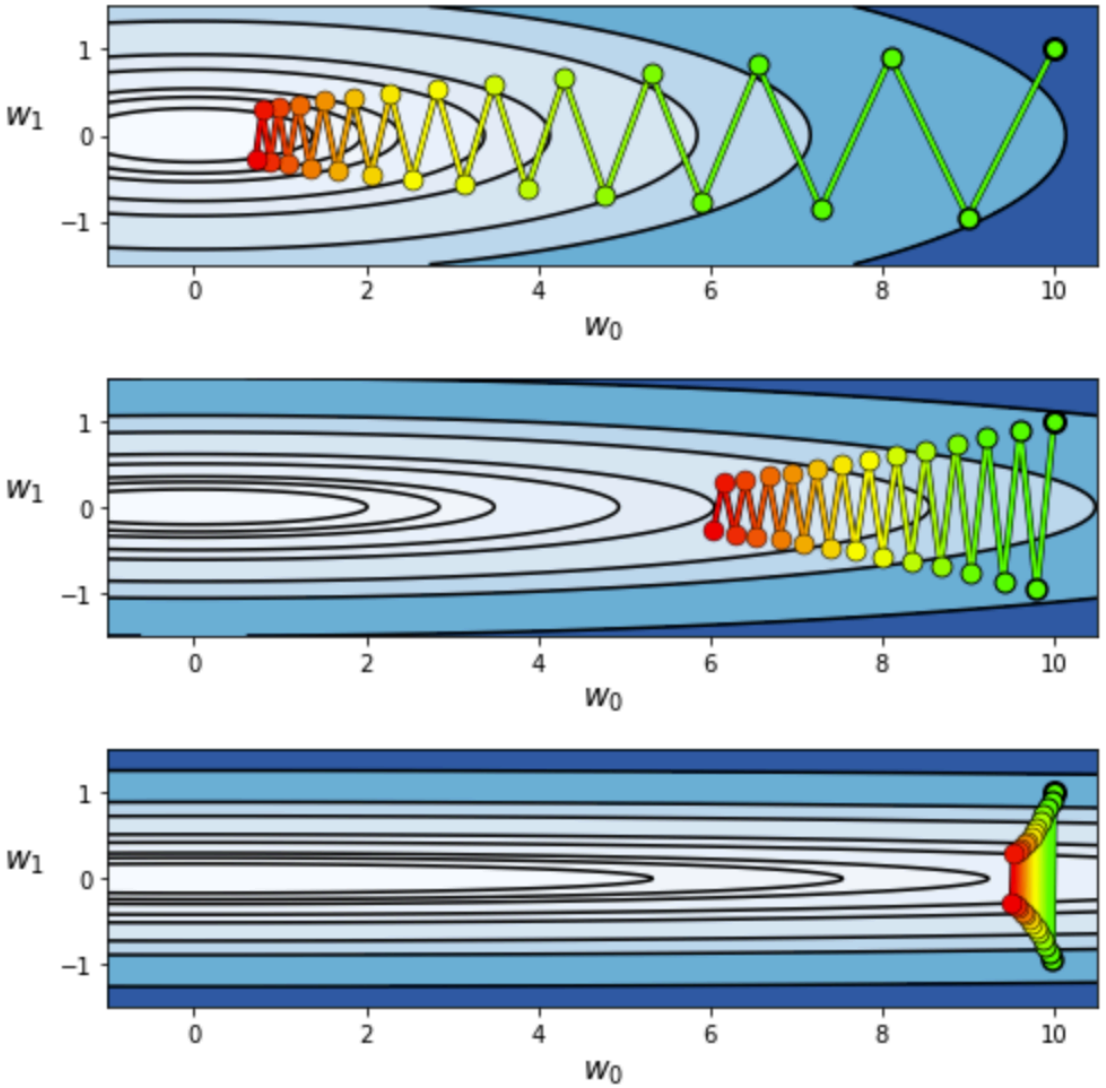
- Local optimization: start with a single input point, sequentially refine it, drive it towards the approximate minimum point.

Local optimization in updating parameters

The general framework is: $\boldsymbol{w}^k = \boldsymbol{w}^{k-1} + \boldsymbol{d}^{k-1}$. $\boldsymbol{d}^{k-1}$: descent direction. Methods differ in how to look for the descent direction:

- Zero-order approach: random search; coordinate search/descent;
- First-order approach: gradient descent - $\boldsymbol{w}^k = \boldsymbol{w}^{k-1} - \alpha \nabla g(\boldsymbol{w}^{k-1})$;
- Second-order approach: newton's method - $\boldsymbol{w}^k = \boldsymbol{w}^{k-1} - (\nabla^2 g(\boldsymbol{w}^{k-1}))^{-1} \nabla g(\boldsymbol{w}^{k-1})$;

Zig-zag phenomenon



The negative gradient direction oscillates rapidly and will slow down the optimization.

This is because the gradient direction is always perpendicular to the contour. When the contour becomes narrow, the gradient direction is nearly parallel to each other. Standardization can help with this.

https://colab.research.google.com/github/jermwatt/machine_learning_refined/blob/main/notes/3_First_order_methods/3_6_Problems.ipynb

Gradient descent with momentum

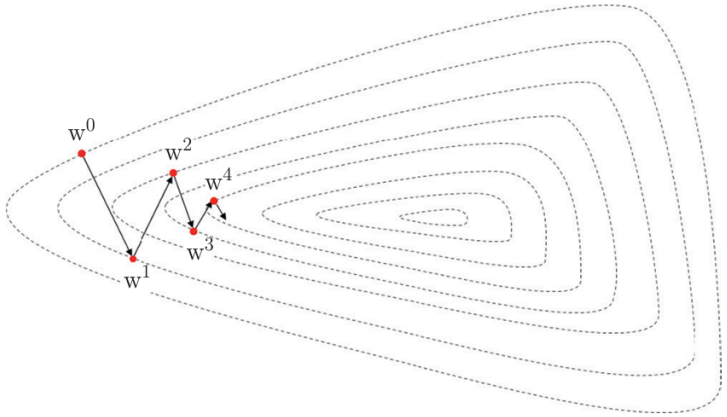


Fig. 7.4 A figurative illustration of gradient steps toward the minimum of a function in two dimensions. Note that the gradient step directions are perpendicular to the contours of the surface shown with dashed ellipses.

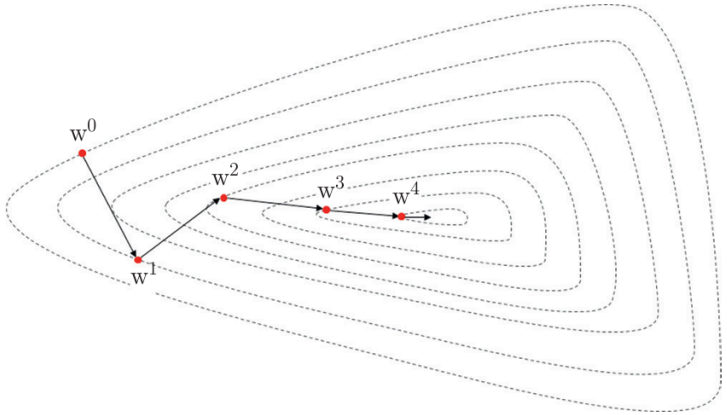


Fig. 7.5 A figurative illustration of momentum-adjusted gradient descent steps toward the minimum of the same function shown in Fig. 7.4. The addition of the momentum term averages out the zig-zagging inherent in standard gradient descent steps.

Adding momentum term to the gradient descent step also helps with the zig-zag phenomenon:

$$\boldsymbol{w}^k = \boldsymbol{w}^{k-1} - \alpha \nabla g\left(\boldsymbol{w}^{k-1}\right) + \beta(\boldsymbol{w}^{k-1} - \boldsymbol{w}^{k-2});$$

Main points about ALE plot

- General goal: evaluate effects of different predictors on the response / give an idea of the importance of predictors for black-box models;

- Problem of PD plot: the expectation in the form $f_{1,PD}(x_1) = E[f(x_1, X_1, \dots, X_k)]$ is taken over the range of marginal distribution of $\boldsymbol{X}_{\setminus j}$;

- Interpretation of first-order ALE plot: observe the range of effect, if is linear/nonlinear;

- Interpretation of second-order ALE plot: observe that in the plot, at different values of X_i, ranges of effect when varying X_j are different or approximately the same;

- Correlation and interaction: correlation is describing how two random variables are linearly related, interaction is describing if the effect that one random variable to the model depends on other random variables.