Machine Intelligence 1 **Exercise Sheet 4**
WS 2016/17, Obermayer/Böhmer due: 17.11.2016

Gradient methods for parameter optimization

Exercise T4.1: Linear neuron (tutorial)

To prepare for the homework, we discuss a simple connectionist neuron with linear output function for a real one-dimensional input and output.

(a) Describe the output function $y(x)$ of the neuron in matrix notation.

(b) Derive gradient and Hesse matrix of the quadratic error function.

(c) Solve the optimization of (b) analytically by using a (generalized) matrix inverse.

(d) Calculate the solution when the objective of (c) includes a "weight decay" regularization term as used in *ridge regression*, i.e.

$$\hat{E}(\underline{\mathbf{w}}) = E(\underline{\mathbf{w}}) + \lambda \sum_i w_i^2 \,.$$

Exercise T4.2: Conjugate gradient (tutorial)

(a) How does the convergence speed of *gradient descent* depend on the learning rate η?

(b) Describe how *line search* speeds up convergence.

(c) What is a *conjugate direction* and how can it improve convergence speed?

(d) What is the maximal number of iterations of *conjugate gradient descent* for a linear neuron in a n dimensional input space, a one-dimensional output and with a quadratic cost function?

Exercise H4.1: Line search (homework, 4 points)

In this exercise you will analyze line search at the simple example of a linear neuron with quadratic cost function. Here we optimize the cost function along a given direction $\underline{\mathbf{d}}_t$:

$$\underline{\mathbf{w}}_{t+1} = \underline{\mathbf{w}}_t - \eta_t \, \underline{\mathbf{d}}_t \,.$$

(a) (1 point) Derive the 2nd order Taylor approximation of an arbitrary $E^T(\underline{\mathbf{w}}_{t+1})$ around $\underline{\mathbf{w}}_t$.

(b) (1 point) Derive a bound on the step size η_t using the above approximation in $E^T_{[\underline{\mathbf{w}}_{t+1}]} \overset{!}{\leq} E^T_{[\underline{\mathbf{w}}_t]}$.

(c) (1 point) Derive the optimal step size η_t^* for cost function $E^T_{[\underline{\mathbf{w}}]} = \frac{1}{2}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)^\top \underline{\mathbf{H}}(\underline{\mathbf{w}} - \underline{\mathbf{w}}^*)$ by minimizing the cost function w.r.t. η. Make sure your solution depends only on known quantities like the weight vector $\underline{\mathbf{w}}_t$, the gradient $\underline{\nabla} E^T_{[\underline{\mathbf{w}}_t]}$ and/or the Hessian $\underline{\mathbf{\Delta}} E^T_{[\underline{\mathbf{w}}_t]}$.

(d) (1 point) Prove that the gradient $\underline{\nabla} E^T_{[\underline{\mathbf{w}}_{t+1}]}$ after one update step with *line search* is orthogonal to the optimized direction $\underline{\mathbf{d}}_t$.

Solution

(a) Let the gradient $\underline{\mathbf{g}}_t := \underline{\nabla} E^T_{[\underline{\mathbf{w}}_t]}$ and the Hesse matrix $\underline{\mathbf{H}}_t := \underline{\boldsymbol{\Delta}} E^T_{[\underline{\mathbf{w}}_t]}$, then

$$\begin{aligned} E^T_{[\underline{\mathbf{w}}_{t+1}]} &\approx E^T_{[\underline{\mathbf{w}}_t]} + (\underline{\mathbf{w}}_{t+1} - \underline{\mathbf{w}}_t)^\top \underline{\mathbf{g}}_t + \tfrac{1}{2}(\underline{\mathbf{w}}_{t+1} - \underline{\mathbf{w}}_t)^\top \underline{\mathbf{H}}_t (\underline{\mathbf{w}}_{t+1} - \underline{\mathbf{w}}_t) \\ &= E^T_{[\underline{\mathbf{w}}_t]} - \eta_t \, \underline{\mathbf{d}}_t^\top \underline{\mathbf{g}}_t + \frac{\eta^2}{2} \underline{\mathbf{d}}_t^\top \underline{\mathbf{H}}_t \underline{\mathbf{d}}_t \,. \end{aligned}$$

(b) $E^T_{[\underline{\mathbf{w}}_{t+1}]} \approx E^T_{[\underline{\mathbf{w}}_t]} - \eta_t \, \underline{\mathbf{d}}_t^\top \underline{\mathbf{g}}_t + \frac{\eta^2}{2} \underline{\mathbf{d}}_t^\top \underline{\mathbf{H}}_t \underline{\mathbf{d}}_t \overset{!}{\leq} E^T_{[\underline{\mathbf{w}}_t]} \quad \Rightarrow \quad \eta_t \overset{!}{\leq} 2 \frac{\underline{\mathbf{d}}_t^\top \underline{\mathbf{g}}_t}{\underline{\mathbf{d}}_t^\top \underline{\mathbf{H}}_t \underline{\mathbf{d}}_t} \,.$

(c) Solve $\min_\eta E^T_{(\underline{\mathbf{w}}_t + 1)}$ by setting the derivative w.r.t. η to zero:

$$\begin{aligned} \frac{\partial E^T_{[\underline{\mathbf{w}}_{t+1}]}}{\partial \eta} &= \left(\frac{\partial E^T_{[\underline{\mathbf{w}}_{t+1}]}}{\partial \underline{\mathbf{w}}_{t+1}} \right)^\top \frac{\partial \underline{\mathbf{w}}_{t+1}}{\partial \eta} = (\underline{\mathbf{H}}_t \underline{\mathbf{w}}_t - \eta_t \underline{\mathbf{H}}_t \underline{\mathbf{d}}_t - \underline{\mathbf{H}}_t \underline{\mathbf{w}}^*)^\top (-\underline{\mathbf{d}}_t) \overset{!}{=} 0 \\ \Rightarrow \quad \eta^* &= \frac{\underline{\mathbf{d}}_t^\top \underline{\mathbf{H}}_t (\underline{\mathbf{w}}_t - \underline{\mathbf{w}}^*)}{\underline{\mathbf{d}}_t^\top \underline{\mathbf{H}}_t \underline{\mathbf{d}}_t} = \frac{\underline{\mathbf{d}}_t^\top \underline{\mathbf{g}}_t}{\underline{\mathbf{d}}_t^\top \underline{\mathbf{H}}_t \underline{\mathbf{d}}_t} \,, \qquad \text{as} \quad \underline{\mathbf{g}}_t = \underline{\mathbf{H}}_t (\underline{\mathbf{w}}_t - \underline{\mathbf{w}}^*) \,. \end{aligned}$$

(d) The gradient is orthogonal to the direction if $\underline{\mathbf{d}}_t^\top \underline{\mathbf{g}}_{t+1} = 0$.

$$\begin{aligned} \underline{\mathbf{g}}_{t+1} &= \underline{\mathbf{H}}_t (\underline{\mathbf{w}}_{t+1} - \underline{\mathbf{w}}^*) = \underline{\mathbf{H}}_t (\underline{\mathbf{w}}_t - \eta_t \underline{\mathbf{d}}_t - \underline{\mathbf{w}}^*) = \underline{\mathbf{g}}_t - \eta_t \underline{\mathbf{H}}_t \underline{\mathbf{d}}_t \\ \underline{\mathbf{d}}_t^\top \underline{\mathbf{g}}_{t+1} &= \underline{\mathbf{d}}_t^\top \underline{\mathbf{g}}_t - \frac{\underline{\mathbf{d}}_t^\top \underline{\mathbf{g}}_t}{\underline{\mathbf{d}}_t^\top \underline{\mathbf{H}}_t \underline{\mathbf{d}}_t} \underline{\mathbf{d}}_t^\top \underline{\mathbf{H}}_t \underline{\mathbf{d}}_t = 0 \end{aligned}$$

Exercise H4.2: Comparison of gradient descent methods (homework, 6 points)

In this exercise we compare the performance of three learning procedures applied to a simple connectionist neuron with linear output function. (i) Gradient (or steepest) descent with constant learning rate, (ii) steepest descent combined with a line search method to determine the learning rate, and (iii) the conjugate gradient method.

Training Data: The training data set consists of three points

$$\{(x^{(\alpha)}, t^{(\alpha)})\} = \{(-1, -0.1), (0.3, 0.5), (2, 0.5)\}\,,$$

i.e. for a given data point, both input and output are scalar values.

Cost function: The gradient for the *quadratic error* function is given by

$$\underline{\mathbf{g}} = \frac{\partial E^T}{\partial \underline{\mathbf{w}}} = \underline{\mathbf{H}}\underline{\mathbf{w}} + \underline{\mathbf{b}}\,, \qquad \text{with} \quad \underline{\mathbf{H}} = \underline{\mathbf{X}}\underline{\mathbf{X}}^T \quad \text{and} \quad \underline{\mathbf{b}} = -\underline{\mathbf{X}}\underline{\mathbf{t}}^T\,.$$

(a) (2 points) *Gradient Descent:* Implement a steepest descent procedure where the weights at iteration $t+1$ are calculated using the weights and the gradient at iteration t

$$\underline{\mathbf{w}}_{t+1} = \underline{\mathbf{w}}_t - \eta \underline{\mathbf{g}}_t,$$

with an adequate learning rate η. Plot the resulting weight vectors from all iterations as a scatter plot (w_0 vs. w_1), and in an additional plot (w_i vs. iterations), to show the development of the parameters during gradient descent.

(b) (2 points) *Line Search:* Implement a line search procedure

$$\underline{\mathbf{w}}_{t+1} = \underline{\mathbf{w}}_t - \eta \underline{\mathbf{g}}_t, \qquad \text{with optimal step size} \qquad \eta = \frac{\underline{\mathbf{g}}_t^T \underline{\mathbf{g}}_t}{\underline{\mathbf{g}}_t^T \underline{\mathbf{H}} \underline{\mathbf{g}}_t}\,.$$

Plot the resulting weight vectors from all iterations as a scatter plot (w_0 vs. w_1), and in an additional plot (w_i vs. iterations), to show the development of the parameters during line search.

(c) (2 points) *Conjugate Gradient:* Implement a conjugate gradient procedure:
Initialize: $\underline{\mathbf{w}}_1, \underline{\mathbf{d}}_1 = -\underline{\mathbf{g}}_1$
while *stopping criterion not satisfied* **do**
minimize E along $\underline{\mathbf{d}}_t$: $\underline{\mathbf{w}}_{t+1} = \underline{\mathbf{w}}_t + \eta_t \underline{\mathbf{d}}_t$ with step size $\eta_t = -\frac{\underline{\mathbf{d}}_t^T \underline{\mathbf{g}}_t}{\underline{\mathbf{d}}_t^T \underline{\mathbf{H}} \underline{\mathbf{d}}_t}$
calculate new gradient $\underline{\mathbf{g}}_{t+1} = \underline{\mathbf{H}}\underline{\mathbf{w}}_{t+1} + \underline{\mathbf{b}}$
calculate new conjugate direction $\underline{\mathbf{d}}_{t+1} = \underline{\mathbf{g}}_{t+1} + \beta_t \underline{\mathbf{d}}_t$ with "momentum"

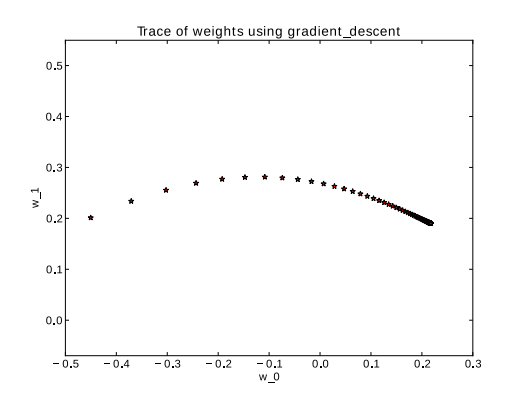
$$\beta_t = -\frac{\underline{\mathbf{g}}_{t+1}^T \underline{\mathbf{g}}_{t+1}}{\underline{\mathbf{g}}_t^T \underline{\mathbf{g}}_t}. \qquad \text{(Fletcher-Reeves form)}$$

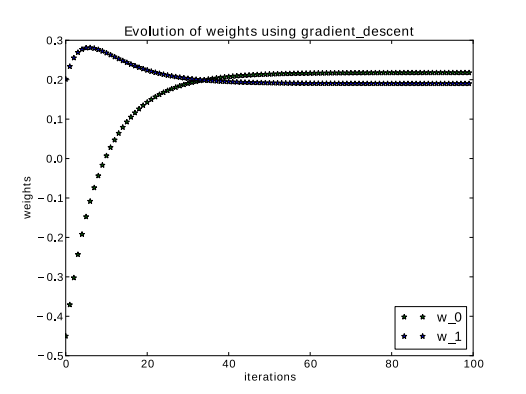
increase $t \leftarrow t+1$
end

Plot the resulting weight vectors from all iterations as a scatter plot (w_0 vs. w_1), and in an additional plot (w_i vs. iterations), to show the development of the parameters during conjugate gradient descent.

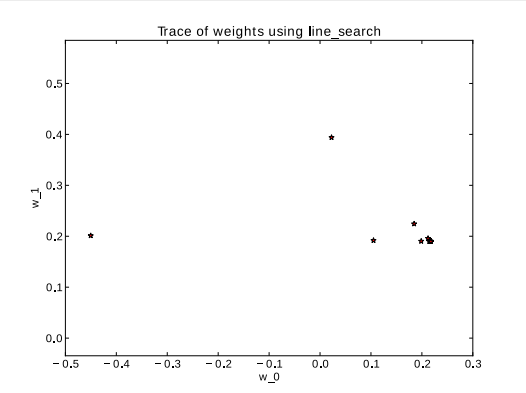
Solution

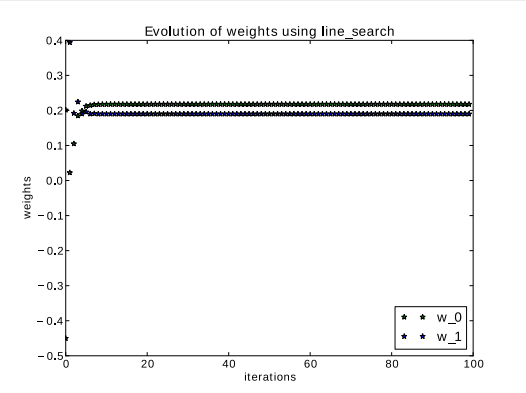
(a) Plot of the weight trace and their evolution over time – fairly slow convergence.



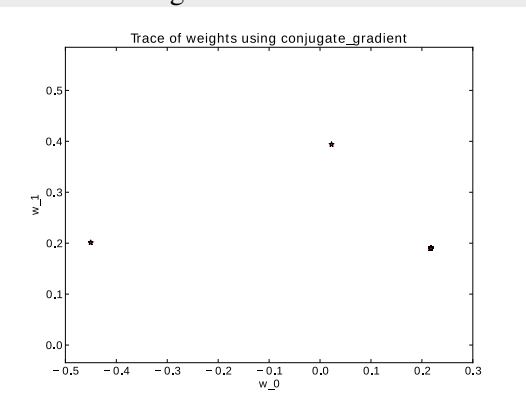


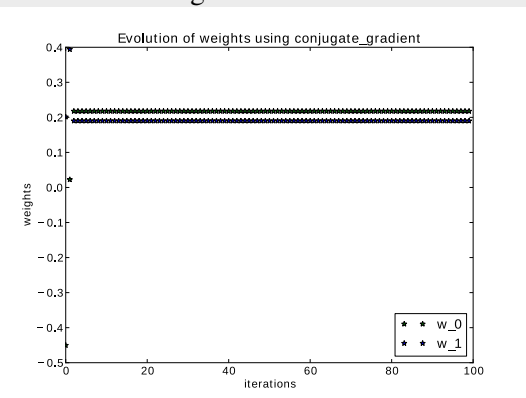
(b) Plot of the weight trace and their evolution over time – much faster convergence in the beginning, noticeable slower close to the minimum.





(c) Plot of the weight trace and their evolution over time – convergence in 2 iterations.





Total 10 points.