EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE FAKULTÄT

Deep Neural Networks

Chapter 4: Numerical Computation

4.1 Overflow and Underflow

- For most real numbers, we incur some approximation error when we represent the number in the computer.
- In many cases, this is just rounding error.
- Underflow occurs when numbers near 0 are rounded to 0.
- We must avoid division by 0, log 0, etc.
- Overflow occurs when numbers with large magnitude are approximated as ∞ or $-\infty$.
- One example of a function that must be stabilized against underflow and overflow is the softmax function:

$$\text{softmax}(\boldsymbol{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}$$

Stabilizing the Softmax Function

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

$$\text{softmax}(\boldsymbol{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}$$

- If for all x_i it holds $x_i = c$, then all of the outputs should be equal to $1/n$.
- Numerically, this may not occur when $|c|$ is large.
- If c is very large and negative, $\exp(c)$ will underflow. The denominator will become 0, so the quotient is undefined.
- When c is very large and positive, $\exp(c)$ will overflow, again resulting in the whole expression being undefined.
- Both difficulties can be resolved by instead evaluating $\text{softmax}(z)$ where $z = x - \max_i x_i$. Adding or subtracting a scalar from z leaves the value of softmax unchanged.
- This results in the largest argument to exp being 0, which rules out the possibility of overflow.
- Likewise, at least one term in the sum has a value of 1, which rules out underflow in the denominator, which would lead to a division by zero.

4.2 Poor Conditioning

- Conditioning refers to how rapidly a function changes with respect to small changes in its inputs. Functions that change rapidly when their inputs are changed slightly can be problematic
- Matrices may be poorly conditioned.
- Consider the function $f(\boldsymbol{x}) = \boldsymbol{A}^{-1}\boldsymbol{x}$. When $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ has an eigenvalue decomposition, its condition number is

$$\max_{i,j} \left| \frac{\lambda_i}{\lambda_j} \right|$$

- This is the ratio of the magnitude of the largest and smallest eigenvalue. When it is large, matrix inversion is particularly sensitive to error in the input.

4.3 Gradient Based Optimization

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- Most deep learning algorithms involve optimization.
- The function we want to minimize or maximize is called the objective function or objective criterion. We may also call it the cost function, loss function, or error function.
- We denote the value x that minimizes or maximizes a function with a superscript x^*. $\boldsymbol{x}^* = \arg\min f(\boldsymbol{x})$
- Suppose we have a function $y = f(x)$, where both x and y are real numbers.
- The derivative of f is denoted as $f'(x)$ or as $\frac{dy}{dx}$. It gives the slope of $f(x)$ at point x.

$$y = f(x+\epsilon) \approx f(x) + \epsilon \cdot f'(x)$$

- We know that $f(x - \epsilon \cdot sign(f'(x))) < f(x)$ for small ϵ.

Gradient Descent Algorithm

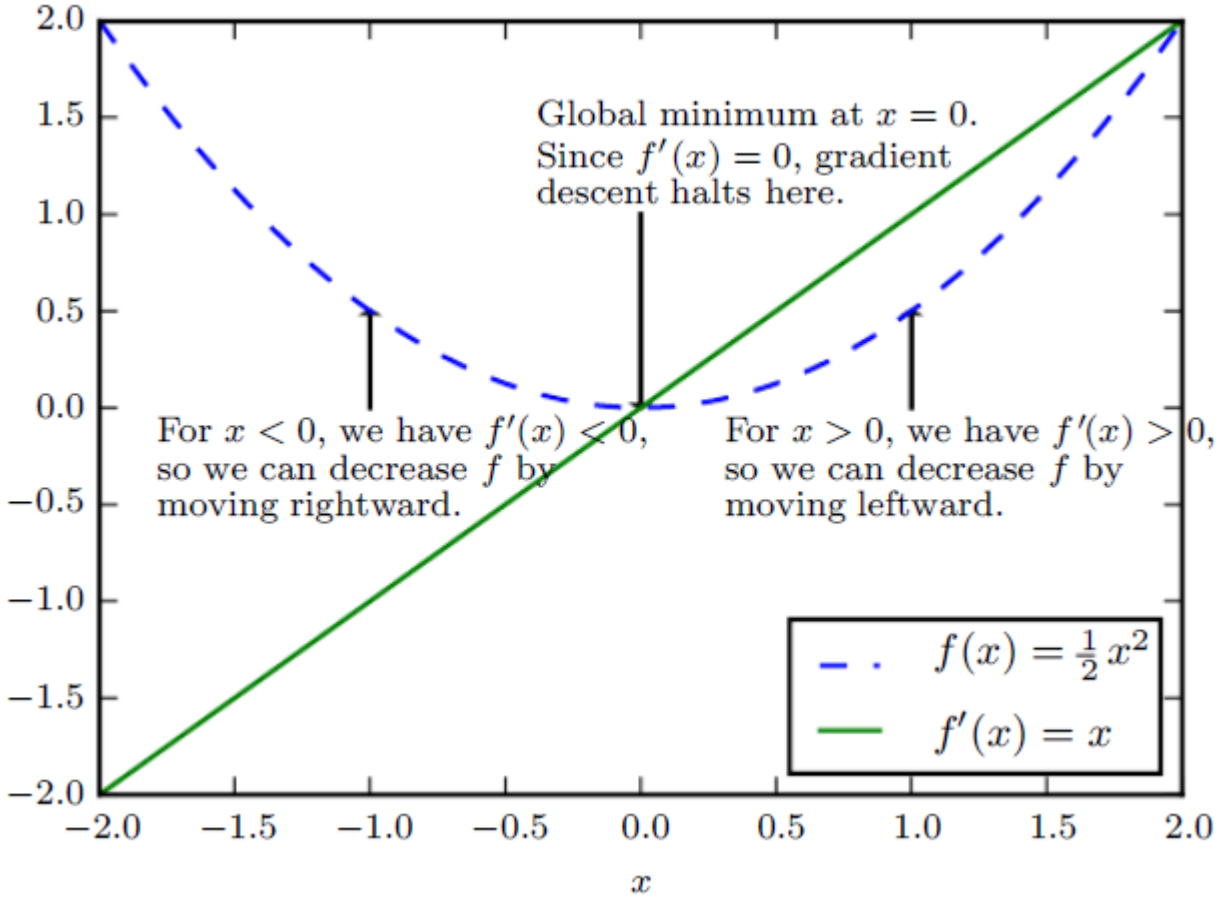

Fig. 4.1: How the gradient descent algorithm uses the derivative of a function to follow the function downhill to a minimum.

Local and Global Minima and Maxima

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- We can thus reduce $f(x)$ by moving in small steps with opposite sign of the derivative. This is called gradient descent.
- Points where $f'(x) = 0$ are called critical points.
- A local minimum is a point where $f(x)$ is lower than all neighboring points.
- A local maximum is a point where $f(x)$ is higher than all neighboring points.
- Some critical points are neither maxima nor minima, but saddle points (see Fig. 4.2).
- A point with the lowest value of $f(x)$ is a global minimum.
- In deep learning we often optimize functions with many local optima.

Critical Points

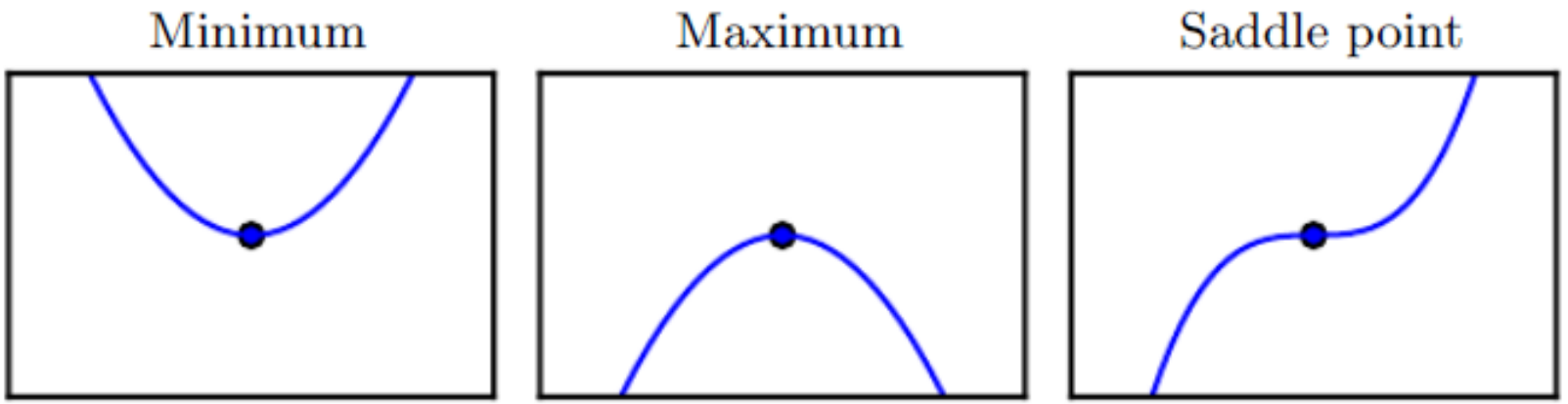

Fig. 4.2: Examples of each of the three types of critical points in 1-D. A critical point is a point with zero slope.
Such a point can either be a local minimum, a local maximum, or a saddle point,

Local Optima of Different Quality

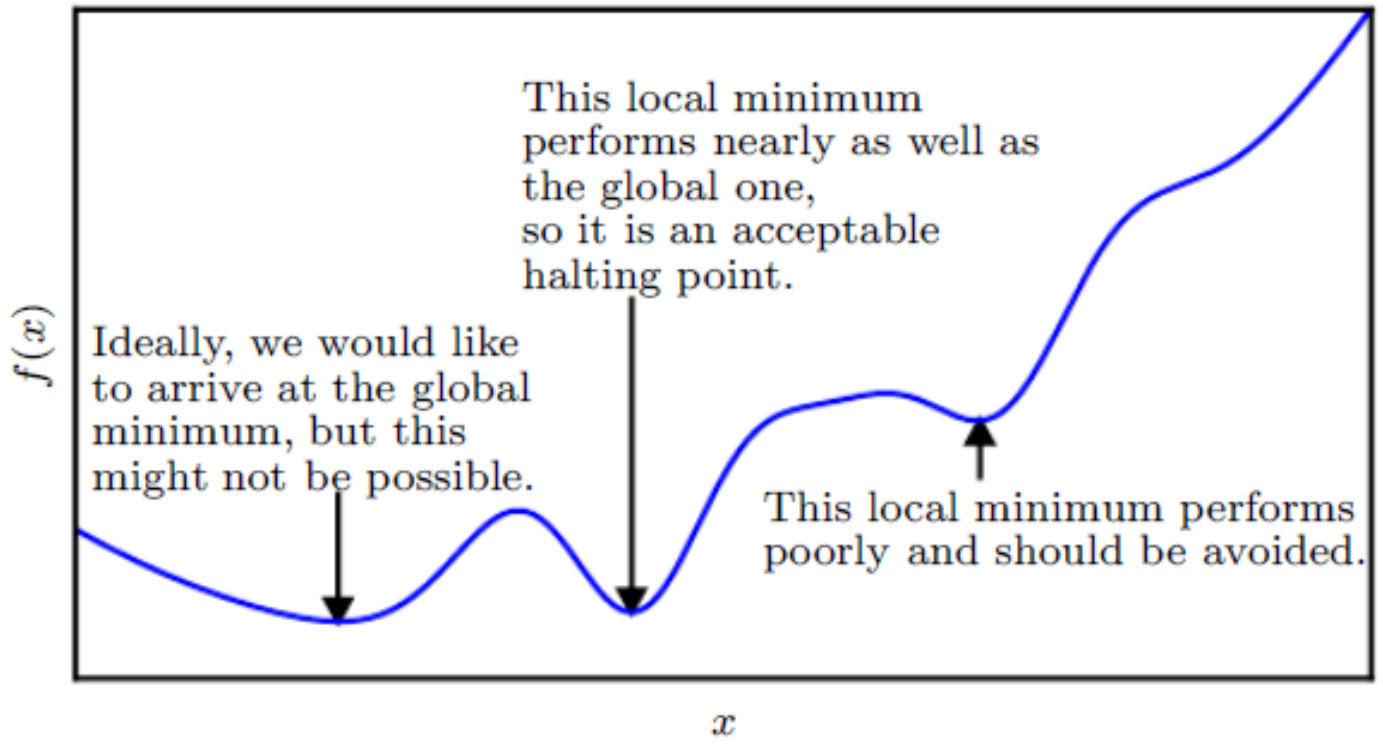

Fig 4.3: Optimization algorithms may fail to find a global minimum when there are multiple local minima or plateaus.

Partial Derivatives

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- The directional derivative in direction $\boldsymbol{u}$ (a unit vector) is the slope of the function f in direction $\boldsymbol{u}$.
- It holds that $\frac{\partial}{\partial \alpha} f(\boldsymbol{x} + \alpha \boldsymbol{u}) = \boldsymbol{u}^{\mathrm{T}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ when $\alpha = 0$.
- To minimize f, we want to find the direction in which f decreases fastest. We use the directional derivative

$$\min_{\boldsymbol{u}, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \boldsymbol{u}^{\mathrm{T}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \min_{\boldsymbol{u}, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \|\boldsymbol{u}\|_2 \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\|_2 \cos\theta$$

  where θ is the angle between $\boldsymbol{u}$ and the gradient.
- Substituting in $\|\boldsymbol{u}\|_2 = 1$ and ignoring factors that do not depend on $\boldsymbol{u}$, this simplifies to $\min_u \cos\theta$.
- This is minimized when $\boldsymbol{u}$ points in the opposite direction of the gradient.

Gradient Descent and Line Search

- In other words the gradient points directly uphill, and the negative gradient points directly downhill.
- We can decrease f by moving in the direction of the negative gradient. This is known as the method of steepest descent or gradient descent.
- Gradient descent proposes a new point

$$\boldsymbol{x}' = \boldsymbol{x} - \epsilon \nabla_{\boldsymbol{x}} f(\boldsymbol{x})$$

  where ϵ is the learning rate (a positive scalar).
- Often ϵ is set to a small constant.
- Another approach is to evaluate $f(\boldsymbol{x} - \epsilon \nabla_{\boldsymbol{x}} f(\boldsymbol{x}))$ for several values of ϵ and choose the one with the smallest objective function value. This is called line search.

4.3.1 Jacobian and Hessian Matrices

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- The matrix containing all partial derivatives of a function whose input and output are both vectors is known as a Jacobian matrix.
- If we have a function $f : \mathbb{R}^m \to \mathbb{R}^n$, then the Jacobian matrix $\boldsymbol{J} \in \mathbb{R}^{n \times m}$ of f is defined such that $J_{i,j} = \frac{\partial}{\partial x_j} f(\boldsymbol{x})_i$.
- We may also be interested in the second derivative.
- For $f : \mathbb{R}^m \to \mathbb{R}^n$, the derivative with respect to x_i of the derivative of f with resp. to x_j is denoted as $\frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x})$
- The second derivative measures the curvature.

Curvature of a Function

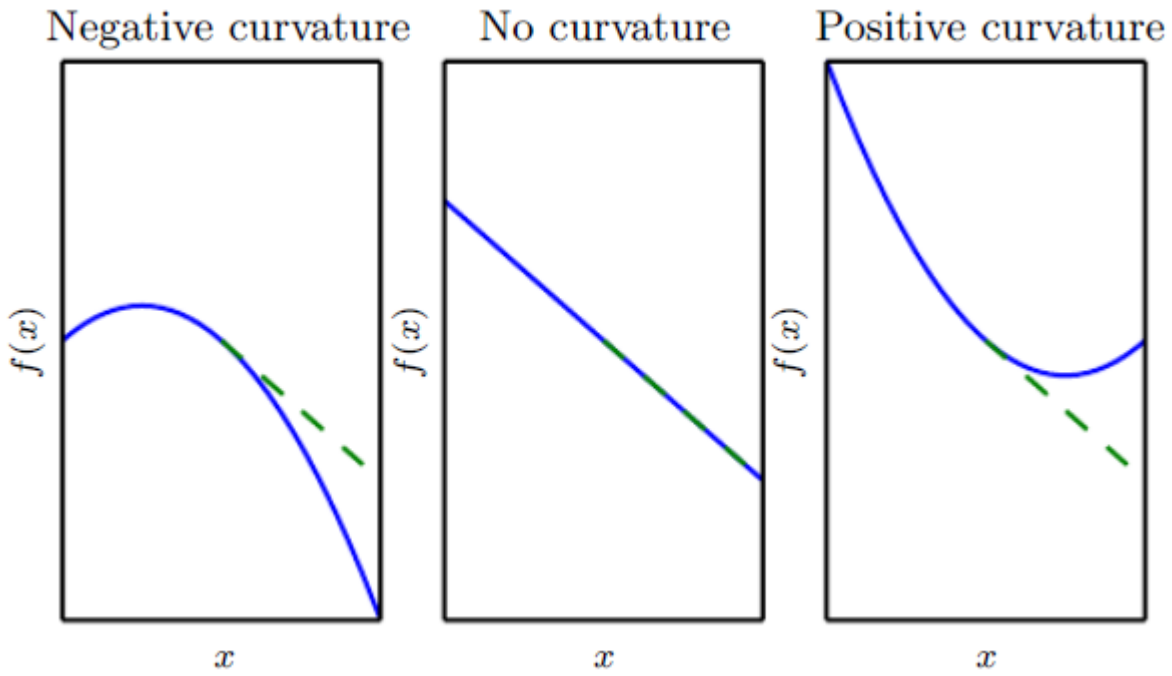

Fig. 4.4: The second derivative determines the curvature of a function. Here we show quadratic functions with various curvature. The dashed line indicates the value of the cost function we would expect based on the gradient information at position x.
At negative curvature, the cost function decreases faster than the gradient predicts.
At no curvature, the gradient predicts the decrease correctly.
At positive curvature, the function decreases slower than expected.

The Hessian Matrix

- When a function f has multiple input dimensions, there are many second derivatives. These derivatives can be collected in a matrix called the Hessian Matrix.

$$\boldsymbol{H}(f)(\boldsymbol{x})_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x})$$

- The Hessian matrix is the Jacobian of the gradient.
- Anywhere that that the second partial derivatives are continuous, the differential operators are commutative:

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x}) = \frac{\partial^2}{\partial x_j \partial x_i} f(\boldsymbol{x})$$

- This implies that $H_{i,j} = H_{j,i}$, so the Hessian matrix is symmetric at such points. It is also real.

The Hessian Matrix

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- As the Hessian matrix is real and symmetric, we can decompose it into a set of real eigenvalues and an orthogonal basis of eigenvectors.
- The second derivative in a specific direction represented by a unit vector $\boldsymbol{d}$ is given by $\boldsymbol{d}^T \boldsymbol{H} \boldsymbol{d}$.
- When $\boldsymbol{d}$ is an eigenvector of $\boldsymbol{H}$, the second derivative in that direction is given by the corresponding eigenvalue.
- For other directions of $\boldsymbol{d}$, the directional 2nd derivative is a weighted average of all of the eigenvalues, with weights between 0 and 1, and eigenvectors that have smaller angle with $\boldsymbol{d}$ receiving more weight. The max. eigenvalue determines the max. 2nd derivative and the min. eigenvalue determines the min. second derivative.

The Hessian Matrix

- We can make a 2nd order Taylor series expansion of the function $f(x)$ around the point $x^{(0)}$:

$$f(x) \approx f(x^{(0)}) + (x - x^{(0)})^{\mathrm{T}} g + \frac{1}{2}(x - x^{(0)})^{\mathrm{T}} H (x - x^{(0)})$$

  where g is the gradient and H is the Hessian at $x^{(0)}$.
- If we use a learning rate of ϵ, then the new point x will be given by $x^{(0)} - \epsilon g$. Substituting this into our formula, we obtain

$$f(x^{(0)} - \epsilon g) \approx f(x^{(0)}) - \epsilon g^{\mathrm{T}} g + \frac{1}{2}\epsilon^2 g^{\mathrm{T}} H g$$

- There are three terms here: the original value of the function, the expected improvement due to the slope of the function, and the correction we must apply to account for the curvature of the function.

The Hessian Matrix

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- When $\boldsymbol{g}^T \boldsymbol{H} \boldsymbol{g}$ is positive, solving for the optimal step size that decreases the Taylor series approximation of the function the most yields

$$\epsilon^* = \frac{\boldsymbol{g}^{\mathrm{T}} \boldsymbol{g}}{\boldsymbol{g}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{g}}$$

- In the worst case, when $\boldsymbol{g}$ aligns with the eigenvector of $\boldsymbol{H}$ corresponding to the max. eigenvalue λ_{max}, then this optimal step size is given by $1/\lambda_{max}$.
- To the extent that the function we minimize can be approximated well by a quadratic function, the eigenvalues of the Hessian thus determine the scale of the learning rate.

$$f(\boldsymbol{x}) = x_1^2 - x_2^2$$

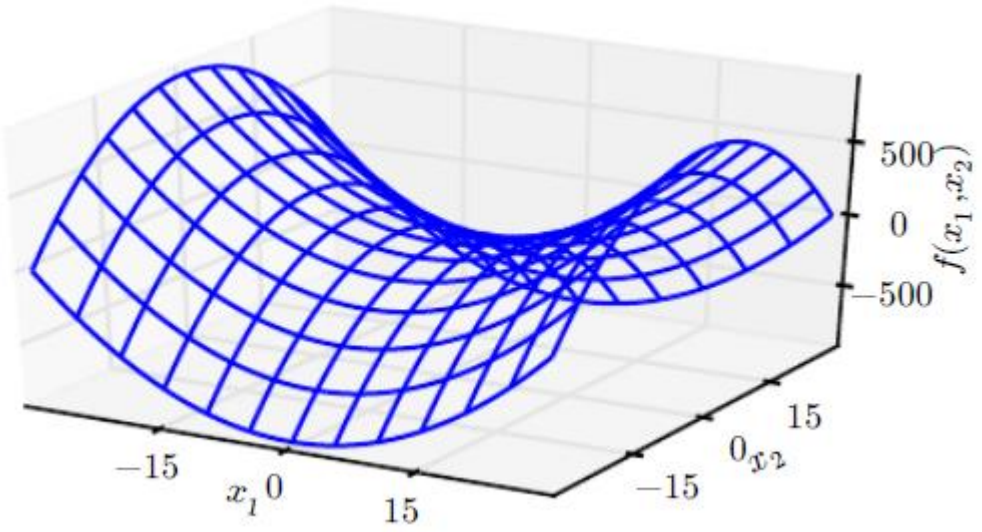

Figure 4.5: A saddle point containing both positive and negative curvature. Along the axis corresponding to x_1, the function curves upward. This axis is an eigenvector of the Hessian with a positive eigenvalue. Along the axis corresponding to x_2, the function curves downward. This direction is an eigenvector of the Hessian with a negative eigenvalue.

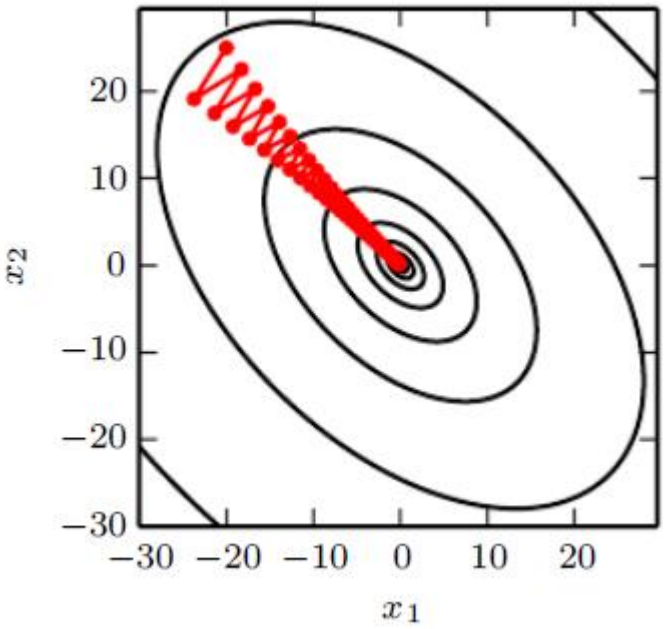

Fig. 4.6: Gradient descent (red line) here fails to exploit the curvature information contained in the Hessian matrix. We minimize a quadratic function $f(\boldsymbol{x})$ whose Hessian has condition number 5. This means that the direction [1, 1] of the largest curvature has five times more curvature than the direction [1, -1] of the smallest curvature. Gradient descent wastes time by repeatedly descending canyon walls (the steepest feature). As the step size is too large, it overshoots and thus needs to descend the opposite wall on the next iteration.

Newton's Method

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- Newton's method uses information from the Hessian matrix to guide the search for a (local) optimum.
- It is based on a second order Taylor series expansion to approximate $f(x)$ near some point $x^{(0)}$:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^{(0)}) + (\boldsymbol{x} - \boldsymbol{x}^{(0)})^{\mathrm{T}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(0)}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(0)})^{\mathrm{T}} \boldsymbol{H}(f)(\boldsymbol{x}^{(0)})(\boldsymbol{x} - \boldsymbol{x}^{(0)})$$

- If we then solve for the critical point of this function, we obtain:

$$\boldsymbol{x}^{*} = \boldsymbol{x}^{(0)} - \boldsymbol{H}(f)(\boldsymbol{x}^{(0)})^{-1} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(0)})$$

- When f is a positive definite quadratic function, Newton's method consists of applying the above equation only once to jump to the minimum of the function directly.

Newton's Method

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- When f is not truly quadratic but locally positive definite quadratic, Newton's method consists of applying the above equation multiple times.
- Iteratively updating the approximation and jumping to the minimum of the approximation can reach the critical point much faster than gradient descent.
- This is a useful property near a local minimum, but it can be a harmful property near a saddle point.
- Optimization algorithms that use only the gradient are called first-order optimization algorithms.
- Optimization algorithms that also use the Hessian matrix, such as Newton's method, are called second-order optimization algorithms.

Lipschitz Continuous Functions

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- In the context of deep learning we sometimes can gain results by restricting ourselves to functions that are either Lipschitz continuous or have Lipschitz continuous derivatives.
- A Lipschitz continuous function is a function f whose rate of change is bound by a Lipschitz constant $\mathcal{L}$:

$$\forall \boldsymbol{x}, \forall \boldsymbol{y}, \left| f(\boldsymbol{x}) - f(\boldsymbol{y}) \right| \leq \mathcal{L} \left\| \boldsymbol{x} - \boldsymbol{y} \right\|_2$$

- This property assures that a small change in the input will have a small change in the output.
- Lipschitz continuity is also a fairly weak constraint, and many optimization problems in deep learning can be made Lipschitz continuous with minor modifications.

4.4 Constrained Optimization

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- Sometimes we wish to find the maximal or minimal value of $f(\boldsymbol{x})$ for values of $\boldsymbol{x}$ in some set $\mathbb{S}$.
- This is known as constrained optimization.
- Points $\boldsymbol{x}$ that lie in the set $\mathbb{S}$ are called feasible points.
- A frequent constraint is a norm constraint, like $\|\boldsymbol{x}\| \leq 1$.
- A sophisticated approach is to design a different, unconstrained optimization problem whose solution can be converted into a solution to the original constrained optimization problem.
- E.g., if we want to minimize $f(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^2$ with $\|\boldsymbol{x}\| = 1$, we can instead minimize $g(\theta) = f([\cos\theta \quad \sin\theta]^T)$ with respect to θ, then return $[\cos\theta, \sin\theta]$ as the solution to the original problem.

Karush-Kuhn-Tucker Approach

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- The Karush–Kuhn–Tucker (KKT) approach provides a very general solution to constrained optimization.
- With the KKT approach, we introduce a new function called the generalized Lagrangian or generalized Lagrange function.
- To define the Lagrangian, we first need to describe $\mathbb{S}$ in terms of equations and inequalities. We describe $\mathbb{S}$ in terms of m functions $g^{(i)}$ and n functions $h^{(j)}$ so that
$$\mathbb{S} = \left\{ \boldsymbol{x} \middle| \forall i, g^{(i)}(\boldsymbol{x}) = 0 \text{ and } \forall j, h^{(j)}(\boldsymbol{x}) \leq 0 \right\} .$$
- The equations with $g^{(i)}$ are called equality constraints.
- The inequalities involving $h^{(j)}$ are called inequality constraints.

Karush-Kuhn-Tucker Approach

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- We introduce new variables λ_i and α_j for each constraint, these are called the KKT multipliers.
- The generalized Lagrangian is then defined as

$$L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\alpha}) = f(\boldsymbol{x}) + \sum_i \lambda_i g^{(i)}(\boldsymbol{x}) + \sum_j \alpha_j h^{(j)}(\boldsymbol{x})$$

- We can now solve a constrained minimization problem using unconstrained optimization of the generalized Lagrangian.
- As long as at least one feasible point exists and $f(\boldsymbol{x}) < \infty$ then

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha},\boldsymbol{\alpha}\geq 0} L(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\alpha})$$

has the same optimal function value and optimal points $\boldsymbol{x}$ as

$$\min_{\boldsymbol{x}\in\mathbb{S}} f(\boldsymbol{x})$$

Karush-Kuhn-Tucker Approach

- This follows because any time the constraints are satisfied,

$$\max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = f(\boldsymbol{x})$$

- while any time a constraint is violated

$$\max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \infty$$

- These properties guarantee that no infeasible point can be optimal and that the optimum within feasible points is unchanged.
- To perform constrained maximization, we can construct the generalized Lagrange function of $-f(\boldsymbol{x})$, which leads to this optimization problem

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq 0} -f(\boldsymbol{x}) + \sum_i \lambda_i g^{(i)}(\boldsymbol{x}) + \sum_j \alpha_j h^{(j)}(\boldsymbol{x})$$

Karush-Kuhn-Tucker Approach

- We may also convert this to a problem with maximization in the outer loop:

$$\max_{x} \min_{\lambda} \min_{\alpha,\alpha\geq 0} f(x) + \sum_i \lambda_i g^{(i)}(x) - \sum_j \alpha_j h^{(j)}(x)$$

  The sign of the term of the equality constraints does not matter, because the optimization is free to choose any sign for each λ_i.
- The inequality constraints are particularly interesting:
- A constraint $h^{(j)}(x)$ is active if $h^{(j)}(x^*) = 0$.
- A point found at convergence remains a stationary point whether or not the inactive constraints are included.
- As an inactive $h^{(j)}$ has negative value, then the solution to $\min_{x} \max_{\lambda} \max_{\alpha,\alpha\geq 0} L(x,\lambda,\alpha)$ will have $\alpha_j = 0$.

Karush-Kuhn-Tucker Approach

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- We can thus observe that at the solution, $\boldsymbol{\alpha} \odot \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$.
- In other words, for all i, we know that at least one of the constraints $\alpha_i \geq 0$ and $h^{(i)}(\boldsymbol{x}) \leq 0$ must be active at the solution.
- We can say that either the solution is on the boundary imposed by the inequality and we must use its KKT multiplier to influence the solution to x, or the inequality has no influence on the solution and we represent this by zeroing out its KKT multiplier.

Karush-Kuhn-Tucker Conditions

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- A simple set of properties describe the optimal points of constrained optimization problems. These properties are called the Karush-Kuhn-Tucker (KKT) conditions.
- They are necessary conditions, but not always sufficient conditions, for a point to be optimal. The conditions are:
  - The gradient of the generalized Lagrangian is zero.
  - All constraints on both x and the KKT multipliers are satisfied.
  - The inequality constraints exhibit “complementary slackness”: $\boldsymbol{\alpha} \odot \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$.

4.5 Example: Linear Least Squares

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- Suppose we want to find the value of x that minimizes

$$f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$$

- First we obtain the gradient:

$$\nabla_x f(\boldsymbol{x}) = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}$$

- We then can follow this gradient downhill, see Alg. 4.1:

Algorithm 4.1 An algorithm to minimize $f(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||_2^2$ with respect to $\boldsymbol{x}$ using gradient descent, starting from an arbitrary value of $\boldsymbol{x}$.

Set the step size (ϵ) and tolerance (δ) to small, positive numbers.
while $||\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^\top\boldsymbol{b}||_2 > \delta$ **do**
  $\boldsymbol{x} \leftarrow \boldsymbol{x} - \epsilon(\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^\top\boldsymbol{b})$
end while

Example: Linear Least Squares

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- One can also solve this problem using Newton's method. In this case, because the true function is quadratic, the quadratic approximation employed by Newton's method is exact, and the algorithm converges to the global minimum in a single step.
- Now suppose we wish to minimize the same function, but subject to the constraint $x^{\mathrm{T}}x \leq 1$. To do so, we introduce the Lagrangian

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} - 1)$$

- We can now solve the problem

$$\min_{\boldsymbol{x}} \max_{\lambda, \lambda \geq 0} L(\boldsymbol{x}, \lambda)$$

Example: Linear Least Squares

- The smallest-norm solution to the unconstrained least squares problem may be found using the Moore-Penrose pseudoinverse: $x = A^{+} b$. If this point is feasible, then it is the solution to the constrained problem.
- Otherwise, we must find a solution where the constraint is active. By differentiating the Lagrangian with respect to x, we obtain the equation

$$A^{\mathrm{T}} A x - A^{\mathrm{T}} b + 2\lambda x = 0 \qquad \text{or equivalently}$$

$$\left(A^{\mathrm{T}} A + 2\lambda I \right) x = A^{\mathrm{T}} b$$

- Solving for x we obtain

$$x = \left(A^{\mathrm{T}} A + 2\lambda I \right)^{-1} A^{\mathrm{T}} b$$

- The magnitude of λ must be chosen that the result obeys the constraint.

Example: Linear Least Squares

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- We can find this value by performing gradient ascent on λ. To do so, observe $$\frac{\partial}{\partial \lambda} L(\boldsymbol{x}, \lambda) = (\boldsymbol{x}^T \boldsymbol{x} - 1)$$
- When the norm of $\boldsymbol{x}$ exceeds 1, this derivative is positive, so to follow the derivative uphill and increase the Lagrangian with respect to λ, we increase λ.
- Because the coefficient on the $\boldsymbol{x}^T \boldsymbol{x}$ penalty has increased, solving the linear equation for $\boldsymbol{x}$ will now yield a solution with smaller norm.
- The process of solving the linear equation and adjusting λ continues until $\boldsymbol{x}$ has the correct norm and the derivative on λ is 0.