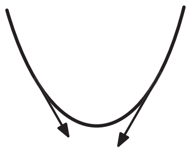
UNCONSTRAINED CONTINUOUS OPTIMIZATION

* Goal: find the minimum of a function $f(x)$.
* Direct way: solve x from $\frac{d}{dx}f(x) = 0$ and check that $\frac{d^2}{dx^2}f(x) > 0$.
* The direct method can be generalized for functions of multiple variables.
* Gradient-based method: start from some tentative solution and move step-by-step towards minimum using derivative.



$$x^{new} \leftarrow x^{old} - \eta \frac{d}{dx} f(x)$$

SINGLE LINEAR NEURON TRAINING (1)

Problem description:

* Consider a single linear neuron with output y and n inputs $x_1, \dots, x_n$. Its behavior is given by:

$$y(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} = \sum_{k=1}^{n} w_k x_k$$

(all inputs are represented by a vector $\boldsymbol{x}$, all weights by a vector $\boldsymbol{w}$).

* The desired behavior is given by input-output pairs $(\boldsymbol{x}^p, d^p)$, $1 \le p \le N$.
* Goal is to find values for the weight parameters that minimize the error:

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{p=1}^{N} (d^p - y(\boldsymbol{x}^p))^2.$$

SINGLE LINEAR NEURON TRAINING (2)

* Solve using gradient search:

$$w_k^{new} \leftarrow w_k^{old} - \eta \frac{\partial}{\partial w_k} E(\boldsymbol{w}).$$

* Iterate until improvements become “small”, e.g.:

$$\sum_{k=1}^{n} \left\| w_k^{new} - w_k^{old} \right\| < \epsilon.$$

* The derivative needed to update the weights:

$$\frac{\partial}{\partial w_k} E(\boldsymbol{w}) = \frac{\partial}{\partial y} E(\boldsymbol{w}) \frac{\partial}{\partial w_k} y = - \sum_{p=1}^{N} (d^p - y(x^p)) x_k^p$$

* The training rule is called the *Widrow-Hoff* learning rule.

METHOD OF LAGRANGE MULTIPLIERS

* Suppose that one wants to know the minimum of a function $F(\boldsymbol{x})$ subject to a constraint $G(\boldsymbol{x}) = 0$.
* Construct a new function $\hat{F}(\boldsymbol{x}, \lambda) = F(\boldsymbol{x}) + \lambda G(\boldsymbol{x})$.
* The solution sought is: $\min_{\boldsymbol{x}, \lambda} \hat{F}(\boldsymbol{x}, \lambda)$.
* Because $\hat{F}(\boldsymbol{x}, \lambda)$ is minimal with respect to λ: $\frac{\partial}{\partial \lambda} \hat{F}(\boldsymbol{x}, \lambda) = G(\boldsymbol{x}) = 0$, the constraint is satisfied; on the other hand, because $G(\boldsymbol{x}) = 0$, any nonzero value can be chosen for λ.

OPTIMAL CONTROL

Problem definition:

* There is a dynamical system with excitation described by:

$$\frac{d}{dt}\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u})$$

* Its initial state $\boldsymbol{x}(0)$ is known.
* The goal is to find time functions for the excitations $\boldsymbol{u}$ that minimize an objective function for the time period $0 < t \le T$.
* The objective function J is composed of two terms, one is a function of the final state $\boldsymbol{x}(T)$ and the other is a function integrated over time:

$$J = \psi[\boldsymbol{x}(T)] + \int_0^T \ell(\boldsymbol{u}, \boldsymbol{x})dt$$

THE EULER-LAGRANGE METHOD (1)

* Provides a solution to the optimal control problem.
* It uses Lagrange multipliers $\lambda(t)$ (a vector of time functions) to construct an objective function $\hat{J}$ that satisfies the dynamical system constraint:

$$\hat{J} = J - \int_{t=0}^{T} \boldsymbol{\lambda}^T[\frac{d}{dt}\boldsymbol{x} - \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u})]dt$$

$$\hat{J} = \psi[\boldsymbol{x}(T)] + \int_0^T \ell(\boldsymbol{u}, \boldsymbol{x})dt - \int_{t=0}^{T} \boldsymbol{\lambda}^T[\frac{d}{dt}\boldsymbol{x} - \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u})]dt$$

THE EULER-LAGRANGE METHOD (2)

* Define the *Hamiltonian* as: $H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{\lambda}^T\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u}) + \ell(\boldsymbol{u}, \boldsymbol{x})$. This leads to the expression for the objective function:

$$\hat{J} = \psi[\boldsymbol{x}(T)] + \int_0^T \left[H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u}) - \boldsymbol{\lambda}^T\frac{d}{dt}\boldsymbol{x}\right]dt$$

* See book for a derivation of the *conditions of optimality*. They are:
  + $-\frac{d}{dt}\boldsymbol{\lambda} = H_x$,

    $$H_x = \begin{bmatrix} \frac{\partial H}{\partial x_1} \\ \vdots \\ \frac{\partial H}{\partial x_n} \end{bmatrix}.$$

  + $\boldsymbol{\lambda}(T) = \psi_x[\boldsymbol{x}(T)]$ (final state condition).
  + Find $\boldsymbol{u}$ such that $H(\boldsymbol{\lambda}, \boldsymbol{x}, \boldsymbol{u})$ is maximized.