Learning for Autonomous Vehicles – Part II

TSFS12: Autonomous Vehicles – Planning, Control, and Learning Systems

Lecture 10-11: Björn Olofsson <bjorn.olofsson@liu.se>

LINKÖPING UNIVERSITY

Purpose of this Lecture

- Provide an **introduction** (focus on usage of methods) to some **core methods in the field of learning for autonomous vehicles**:
  - **Neural networks**,
  - **Reinforcement learning**.

Expected Take-Aways from this Lecture

- Be familiar with how **neural networks** can be used for **learning**.
- Have **basic knowledge** about the formalism of **Markov decision processes** and basic methods for **solving reinforcement-learning problems** in discrete time and finite state and action spaces.
- Be familiar with how **neural networks and reinforcement learning can be combined** to solve problems where the state spaces are **continuous**.

Literature Reading

The following book and article sections are the main reading material for this lecture. References to further reading are provided throughout the slides and at the end of the lecture slides.

- Sections 11.2-11.8 in Hastie, T., R. Tibshirani, J. Friedman, & J. Franklin: *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd Edition, Springer, 2005.
- Sections 1, 4.1-4.4, and 6.1-6.5 in Sutton, R. S., & A. G. Barto: *Reinforcement learning: An introduction*. MIT Press, 2018.
- Scan the content of Mnih, V., Kavukcuoglu, K., Silver, D. et al: ”Human-level control through deep reinforcement learning”, Nature 518, 529–533, 2015.

Outline of the Lecture

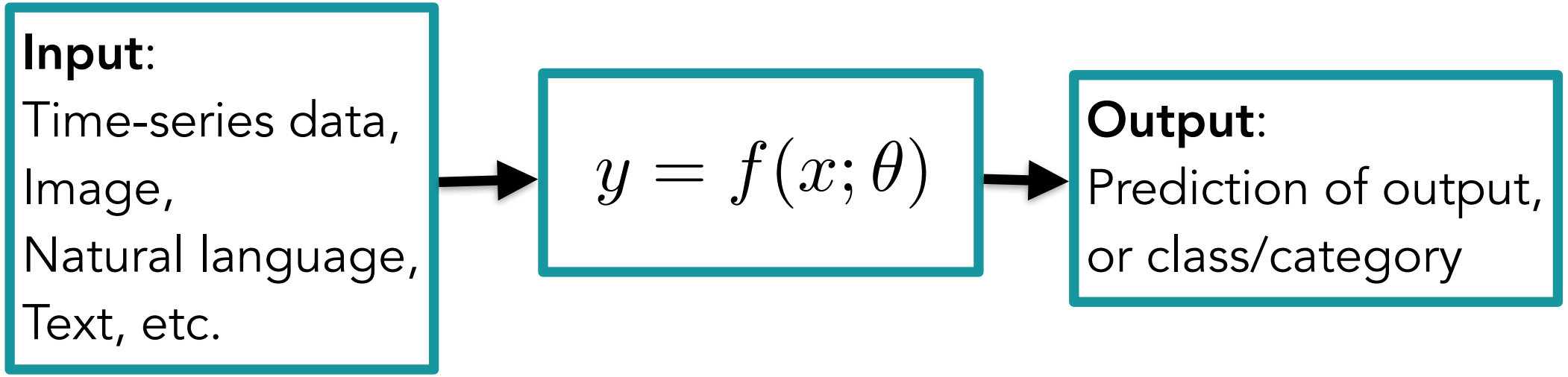
- Learning using **neural networks**.
- Introduction to **reinforcement learning**.
- **Combining** neural networks and reinforcement learning.
- Some **software libraries** for machine learning.

Learning Using Neural Networks

Introduction and Background (1/2)

- Historically: **Inspiration from the human brain**, with its neurons and synapses (connections); activation of a neuron by a signal that reaches a certain threshold.
- A neural network is a **nonlinear function approximator** (often with multiple inputs and outputs), curve-fitting in **high-dimensional spaces**.
- Concept has been described **several decades ago**, and since ~10 years a very **active research area again** because of much **available data**, **efficient algorithms**, and **efficient computational hardware platforms**.

Introduction and Background (2/2)



- Function **parameterized** using parameters in θ (often high-dimensional vector, with thousand, millions, or even billions of elements).
- Often **large amount of training data** $X = (x_1, \ldots, x_N)$ with large N to compute the parameters.

A Nonlinear Function Approximator

- A **complex nonlinear relation** between inputs and outputs can be approximated using a **parametric function**.
- The parameters of the model are determined by **fitting the parameters** to the **training data**.
  - The model uses **specific characteristics** (often called **features**) of the training data, ideally being informative and independent of each other (e.g., specific measured or observed quantities).
- **Separate validation data** are then used to evaluate the fit of the model.

Structure of a Single Hidden-Layer Neural Network

- A neural network consists of **hidden layers** and an **output layer**.
- The basic version contains a single hidden layer.
- A (nonlinear) **activation function** acts on an affine combination of the inputs.

$$Z_m = \sigma(\alpha_{0,m} + \alpha_m^{\mathrm{T}} X), \quad m = 1, \ldots, M,$$
$$T_k = \beta_{0,k} + \beta_k^{\mathrm{T}} Z, \quad k = 1, \ldots, K,$$
$$f_k(X) = g_k(T), \quad k = 1, \ldots, K$$

X – input data
$f_k(X)$ – model output for input X
$\sigma(\cdot)$ – activation function
α, β – parameters
$Z = (Z_1, Z_2, \ldots, Z_M)$
$T = (T_1, T_2, \ldots, T_K)$
$g_k(\cdot)$ – output function

Hastie, T., R. Tibshirani, J. Friedman, & J. Franklin: The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edition, Springer, 2005.

Some Example Activation Functions

- **Linear combination** of the inputs.
- **Sigmoid** (approximation of a step function):

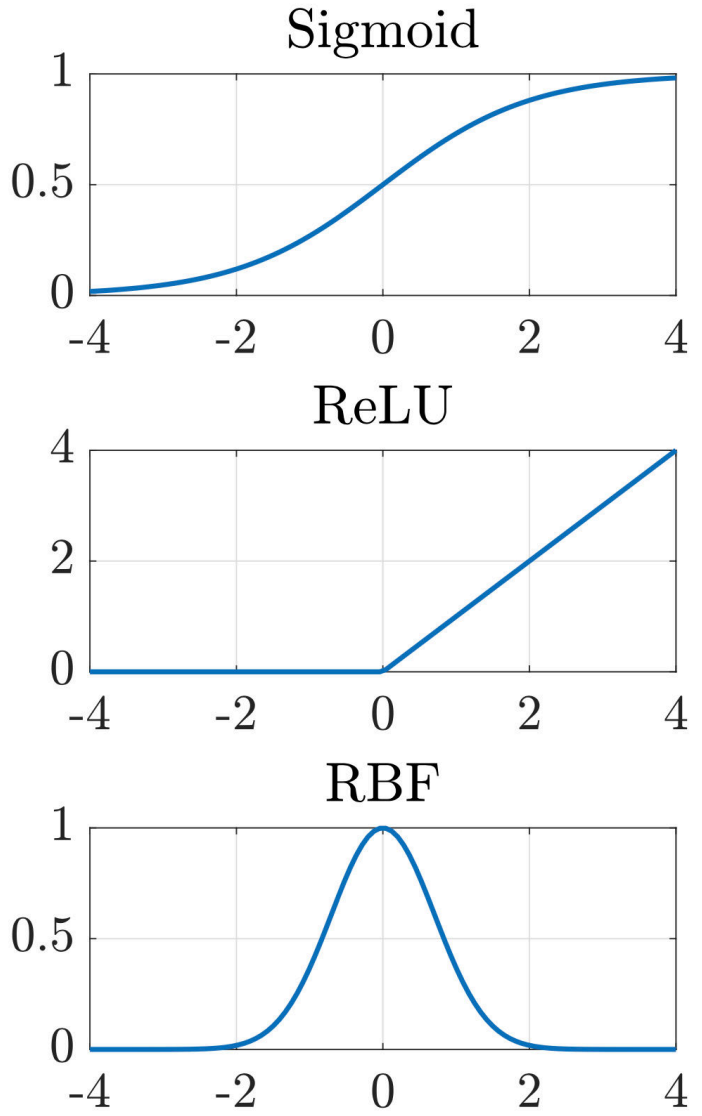
$$\sigma(v) = \frac{1}{1 + \exp(-v)}$$

- Rectifier (Rectified Linear Unit – **ReLU**):

$$\sigma(v) = \max(0, v)$$

- Gaussian radial basis function (**RBF**):

$$\sigma(v) = \exp\left(-\gamma||v - c||^2\right)$$



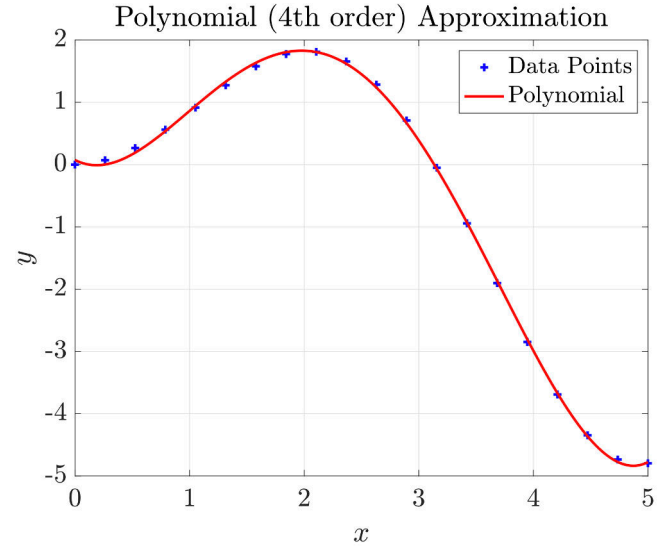
Goodfellow, I., Y. Bengio, & A. Courville: Deep Learning. MIT Press, 2016.

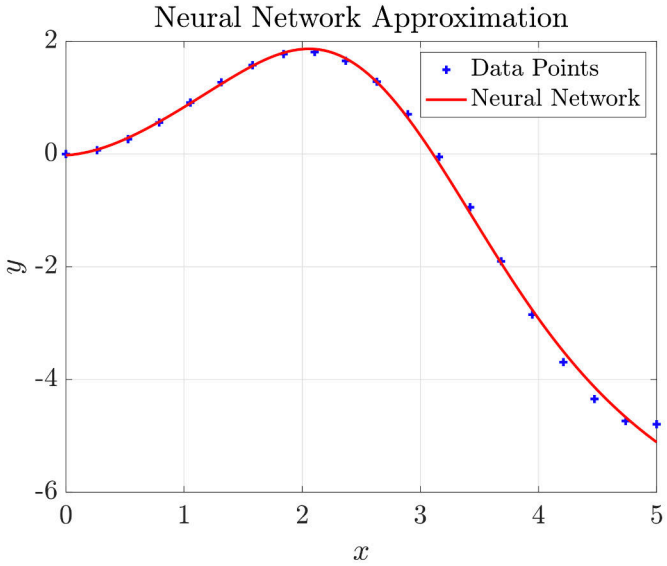
Choices of Output Function (1/2)

- For **regression problems** (i.e., finding relations between dependent and independent variables), the output function can be linear

$$g_k(T) = T_k$$

- Compare with **least-squares approximation** of data points using **polynomials** from previous courses.





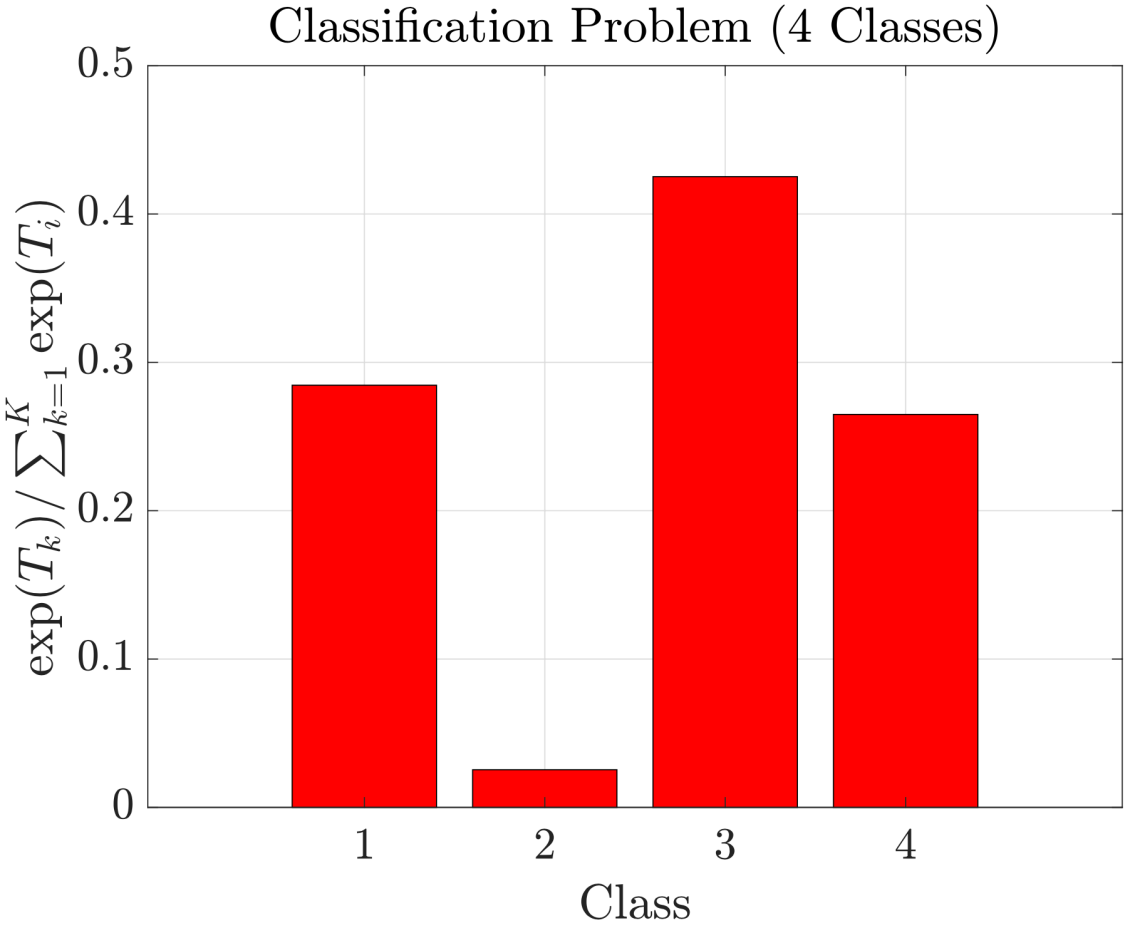
Hastie, T., R. Tibshirani, J. Friedman, & J. Franklin: The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edition, Springer, 2005.

Choices of Output Function (2/2)

- For **classification problems**, the output function can be

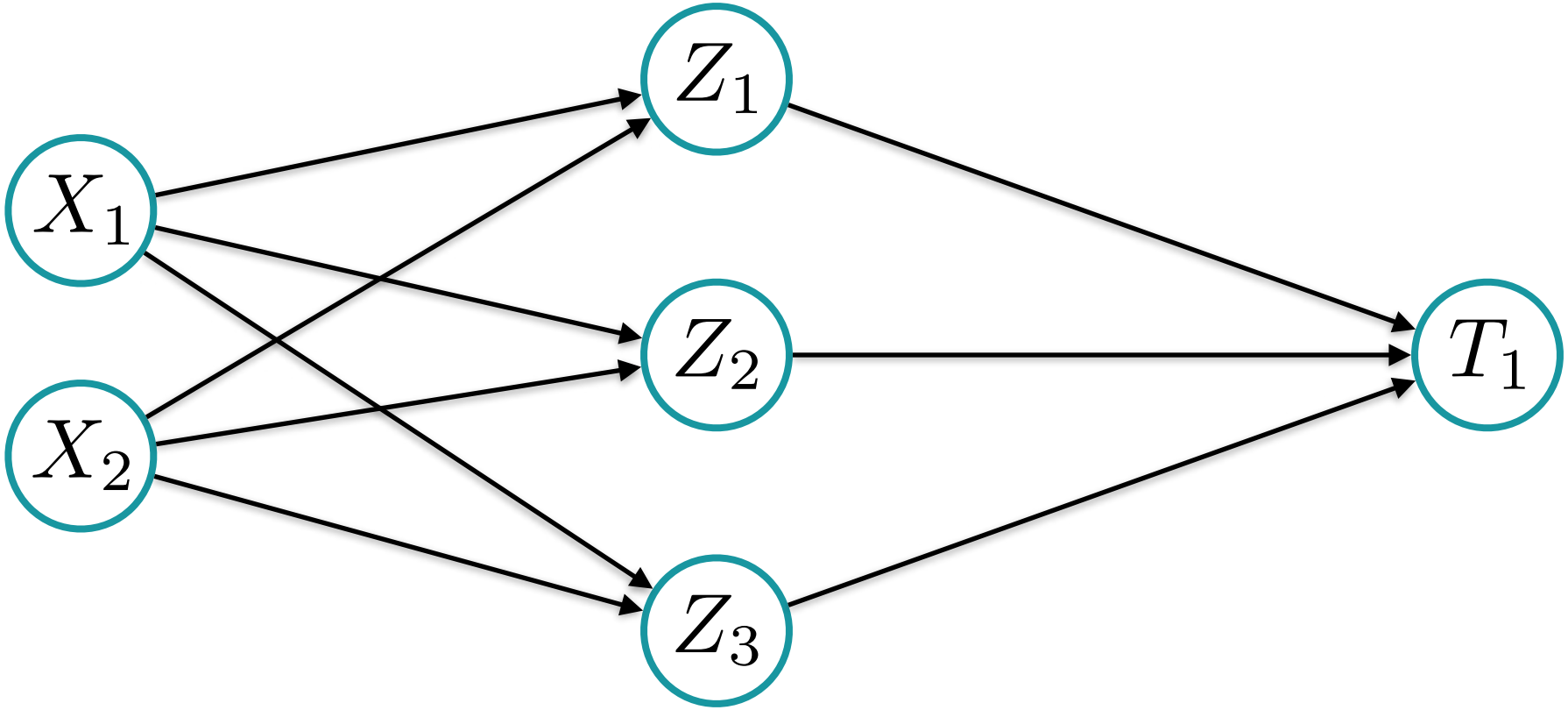
$$g_k(T) = \frac{\exp(T_k)}{\sum_{i=1}^{K} \exp(T_i)}$$

- Consider **example** with **dimension four** in the plot to the right.



Hastie, T., R. Tibshirani, J. Friedman, & J. Franklin: The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edition, Springer, 2005.

Example of a Neural Network



$$\sigma(\alpha_{0,m} + \alpha_m^{\mathrm{T}} X) \qquad \beta_{0,k} + \beta_k^{\mathrm{T}} Z$$

Training of a Neural Network

- Use **training data** to **determine the parameters** θ of the model (e.g., α, β in the example on the previous slide).
- The parameters can be computed by **minimizing a cost function** – an **optimization** problem.
- Examples of cost functions.

Squared-error cost function (regression):

$$J(\theta) = \sum_{i,k} \left(y_{i,k} - f_k(x_i)\right)^2$$

Cross-entropy cost function (classification):

$$J(\theta) = -\sum_{i,k} y_{i,k} \log\left(f_k(x_i)\right)$$

Hastie, T., R. Tibshirani, J. Friedman, & J. Franklin: The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edition, Springer, 2005.

Gradient Descent (1/2)

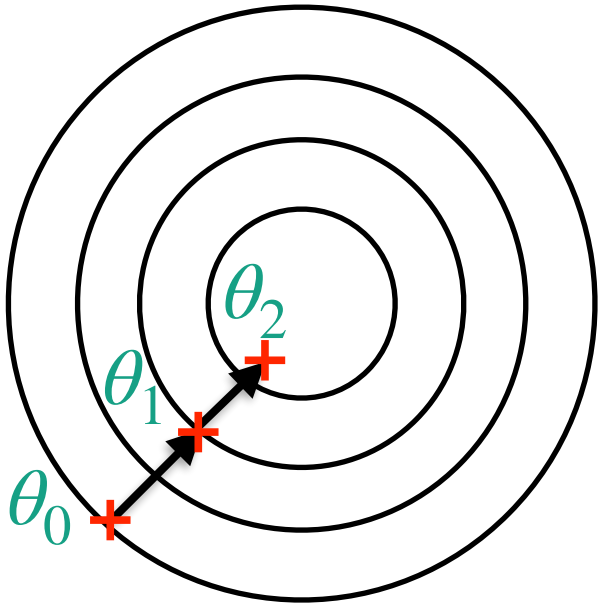
- **Gradient descent** is a method to find a **local minimum** to a (differentiable) cost function using only first-order derivatives.
- Recall from the courses in calculus that a function decreases most rapidly when **moving in the negative direction of the gradient**.

Gradient Descent (2/2)

- **Iterative method** for moving towards a local minimum:

$$\theta_{i+1} = \theta_i - \gamma \nabla J(\theta_i)$$

γ – learning rate (pre-determined parameter)
θ_0 – initial guess for parameters



Stochastic Gradient Descent (SGD)

- Computation of the gradient becomes a challenge when the **number of data points increases to large numbers** or **network output time-consuming to compute**.
- **Stochastic gradient descent** has been suggested to mitigate this:
  - Randomly **select a batch of the data points** (in the limit only one) and perform gradient descent. Idea for **reducing size of parameter-optimization problem**:

$$\sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 \Longrightarrow \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2, \quad n \ll N$$

  - Repeat until a local minimum has been reached (**termination condition**).

Goodfellow, I., Y. Bengio, & A. Courville: Deep Learning. MIT Press, 2016.

Backpropagation

- Recall the **automatic differentiation** method for computing derivatives from Lecture 5.
- Neural networks are often trained using **backpropagation**, which also utilizes the **chain rule** to **compute the desired derivatives** efficiently and accurately (reverse-mode automatic differentiation, see Lecture 5).

Goodfellow, I., Y. Bengio, & A. Courville: Deep Learning. MIT Press, 2016.

Deep Neural Networks

- The neural networks considered so far only consist of one hidden layer.
- **Deep neural networks** comprise multiple hidden layers, where each layer can be considered as an abstraction.
  - Idea: **Features** in the **training data** captured by the neural network.
- **Large variety of network structures** (number of layers, how many neurons in each, types of neurons, connectivity, etc.).
- Typically a **composition of layers** of different network primitives.

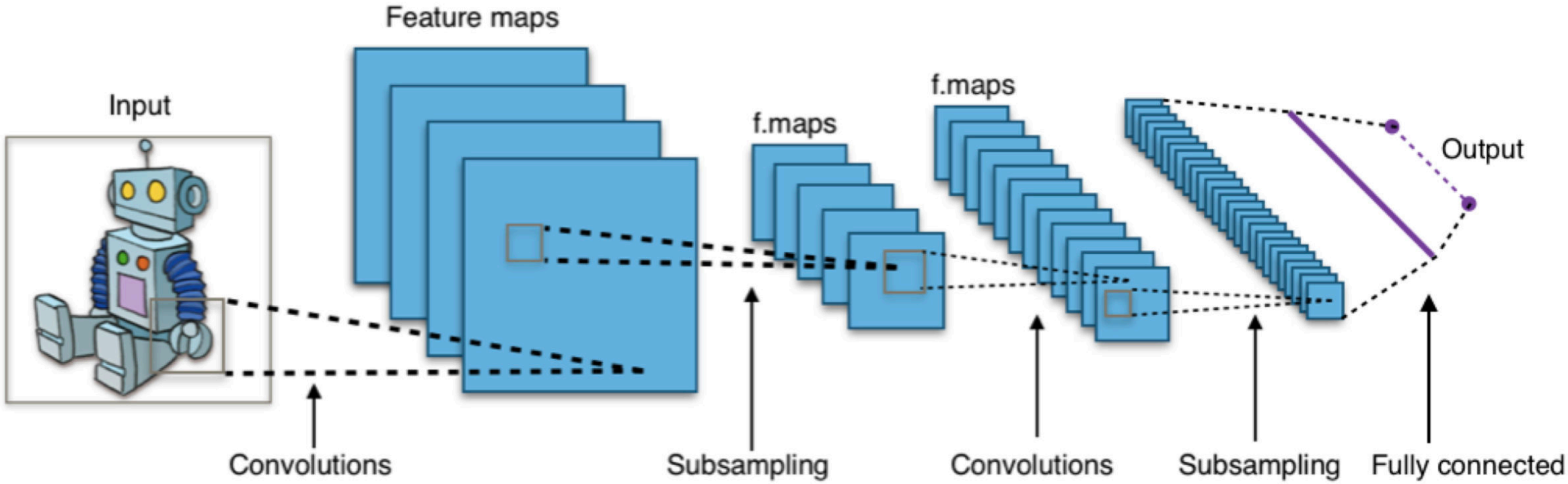
Goodfellow, I., Y. Bengio, & A. Courville: Deep Learning. MIT Press, 2016.

Recurrent and Convolutional Neural Networks

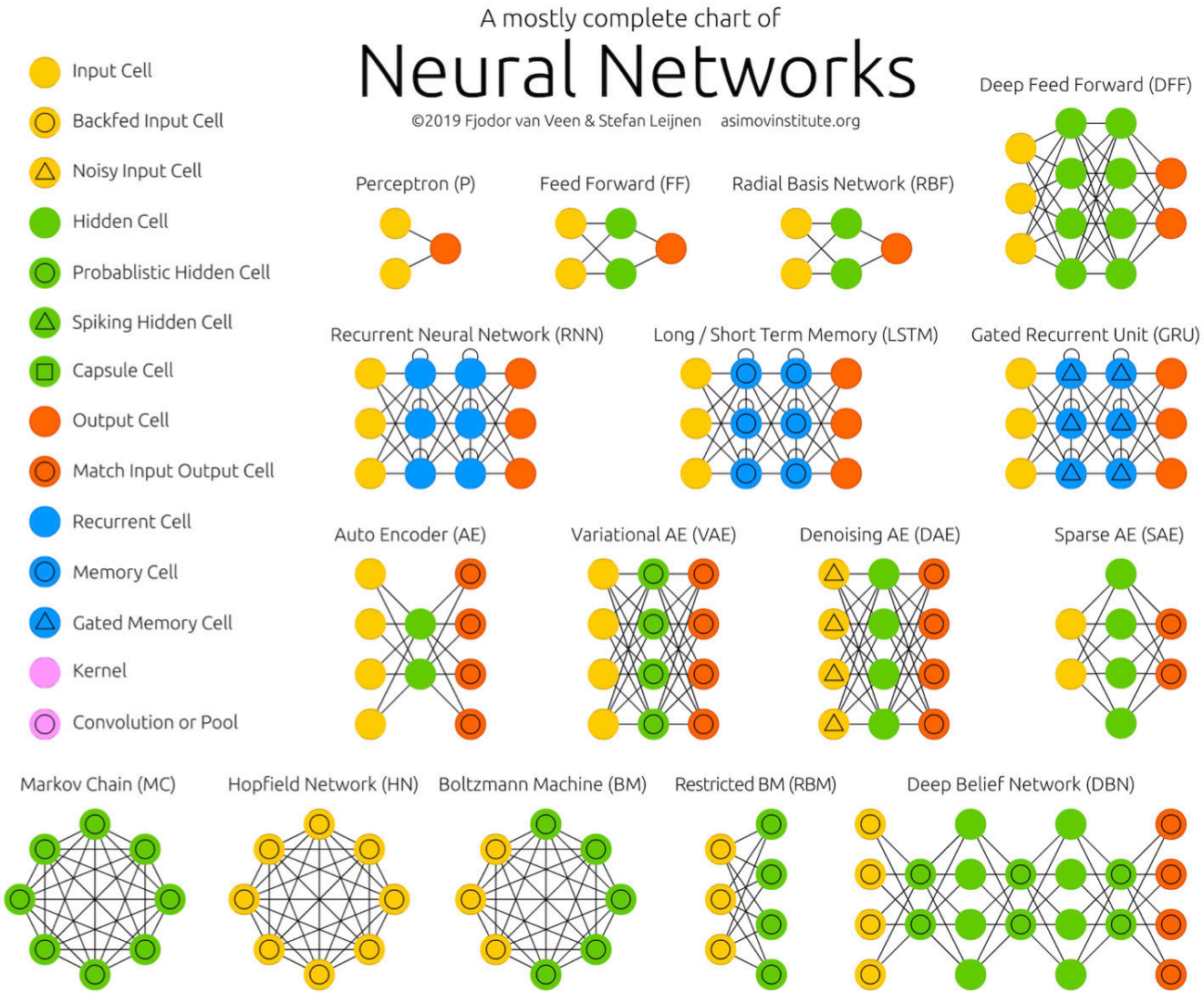
- A **recurrent neural network** also has internal feedback loops (modeling a local memory).
- A **convolutional neural network** comprises layers where convolution operators act.
  - Instead of fully connected layers, **local spatial information** is modeled through the convolutions.
  - Often used for neural networks involving **inputs** with a **grid structure**, like a camera image or written text.

Goodfellow, I., Y. Bengio, & A. Courville: Deep Learning. MIT Press, 2016.

Example Structure of Convolutional Neural Networks

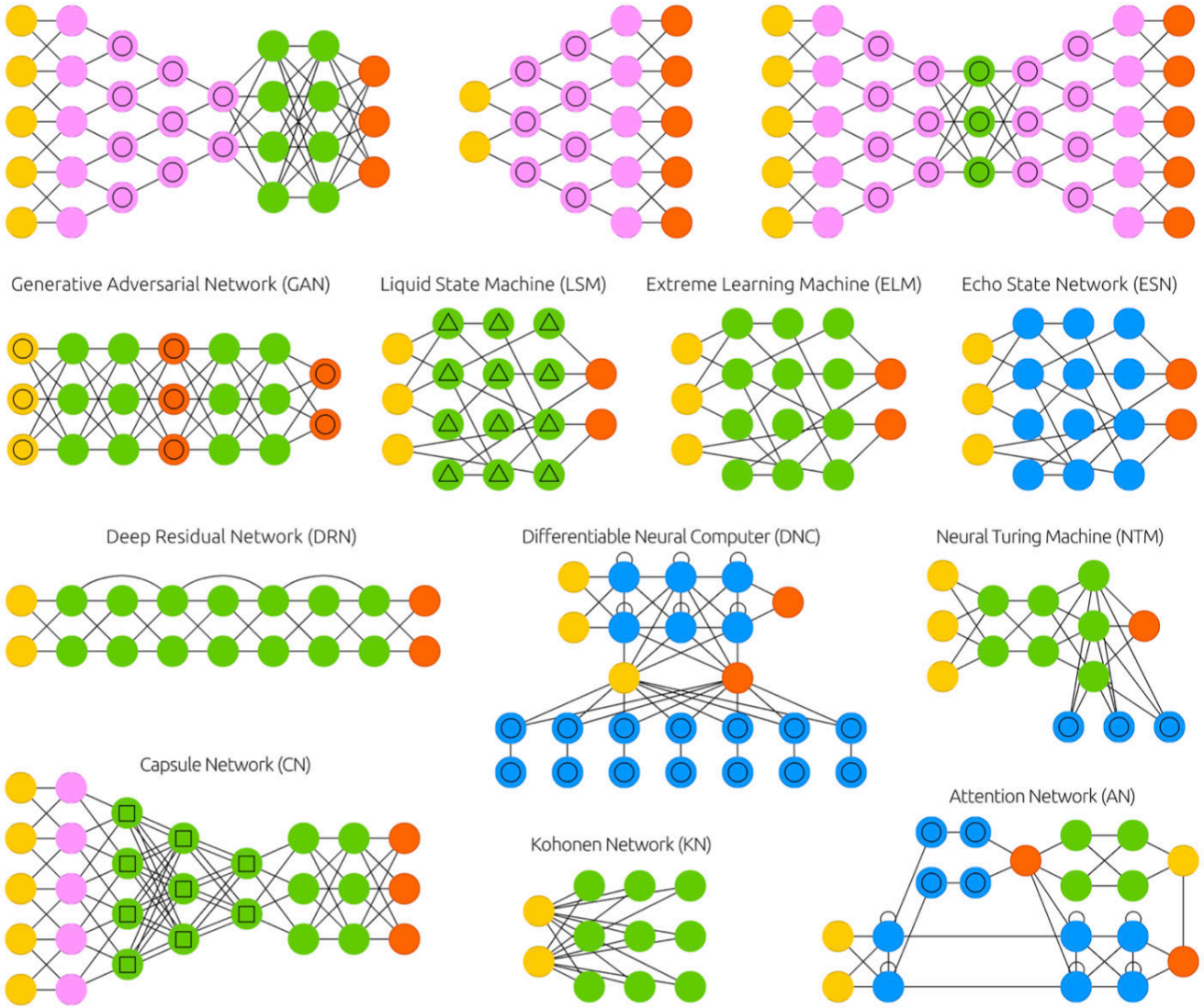


Architectures for Neural Networks (1/2)



Picture from https://www.asimovinstitute.org/neural-network-zoo/

Architectures for Neural Networks (2/2)



Picture from https://www.asimovinstitute.org/neural-network-zoo/

Parameter Optimization

- Several **extensions and variants of stochastic gradient descent** exist for improved performance (e.g., momentum acceleration, Nesterov accelerated gradient, AdaGrad, RMSProp, and Adam).
- Choice of **initial values** of **model parameters** in the gradient descent.

Goodfellow, I., Y. Bengio, & A. Courville: Deep Learning. MIT Press, 2016.

Overfitting of Model Parameters (1/3)

- **Overfitting** of model parameters to training data is a common challenge in any function approximation.
- With **high-dimensional parameter vectors**, it is easy to obtain overfitting to the particular data used for training.
  - For example, being sensitive to **noise in the data** and not modeling the **actual underlying function characteristics**.

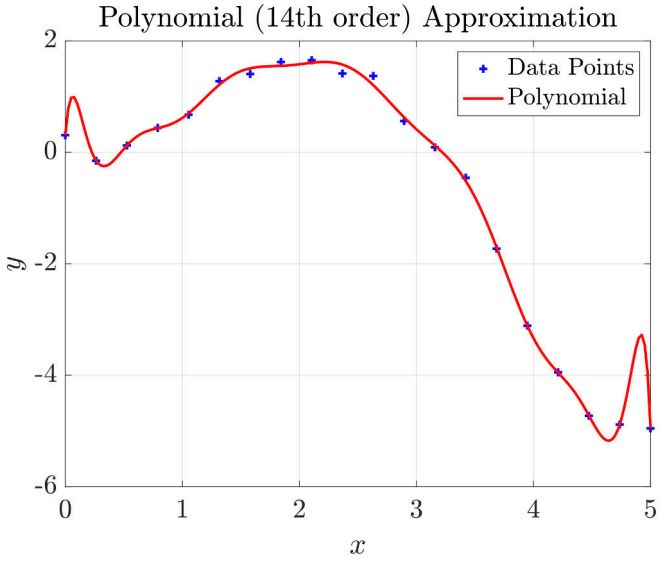
Hastie, T., R. Tibshirani, J. Friedman, & J. Franklin: The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edition, Springer, 2005.
Goodfellow, I., Y. Bengio, & A. Courville: Deep Learning. MIT Press, 2016.

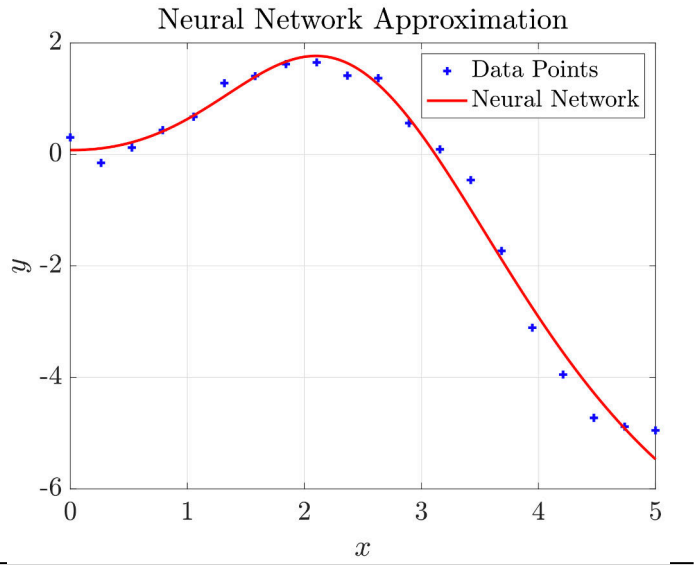
Overfitting of Model Parameters (2/3)

- Example: Fit a high-order **polynomial model** (14th order) and a **neural network model** (30 hidden neurons), respectively.
- 20 data points, where **noise** from a Normal distribution with standard deviation 0.2 has been added, from the function

$$x\sin(x),\ \ x \in [0, 5]$$

- Clear **overfitting** in polynomial model.





Overfitting of Model Parameters (3/3)

- Methods used for mitigating overfitting:
  - **Early termination** in the SGD.
  - **Regularization** terms in cost function, e.g., $\lambda \sum_i \theta_i^2, \ \lambda \text{ weight parameter}$
- Empirical methods to avoid overfitting are, e.g.,
  - **Dropout** (randomly disconnect a subset of the nodes in each phase of the training and then re-connect them again).
- To avoid getting stuck in **local minima** in the optimization, training the network with many different initial guesses of the parameters in SGD is beneficial.

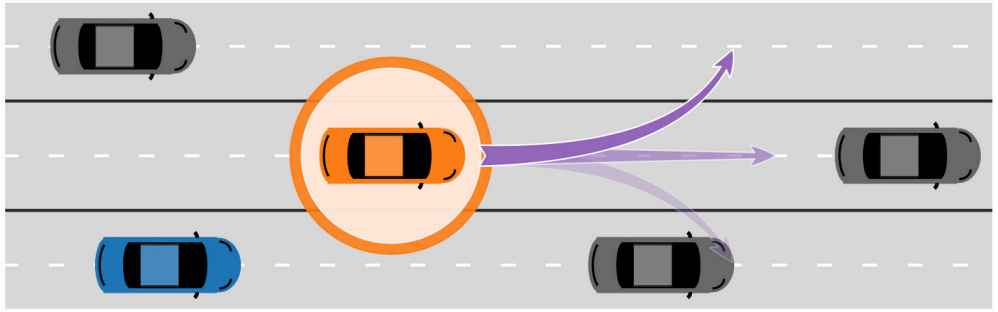
Hastie, T., R. Tibshirani, J. Friedman, & J. Franklin: The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edition, Springer, 2005.
Goodfellow, I., Y. Bengio, & A. Courville: Deep Learning. MIT Press, 2016.

Imbalanced Learning

- **Imbalance in the training data** for a neural network aimed at classification can result in a **biased classifier**.
  - Example is when data from one or more classes are overrepresented compared to other classes.
- One approach to remedy this is to **weight underrepresented classes** when creating the training data set by **sampling** (*with replacement*) from the actual data set.

Imbalanced Learning – Example

- Example: Assume that the task is to **predict if a certain vehicle will change lane** within the next few seconds.
- If the **training data** consist of 99 % cases driving forward, a model always predicting moving forward would have an accuracy of 99 % for that data set.
- Clearly such a predictor **is not satisfactory,** since the important **lane-change situations** will never be predicted – **imbalance in data**.



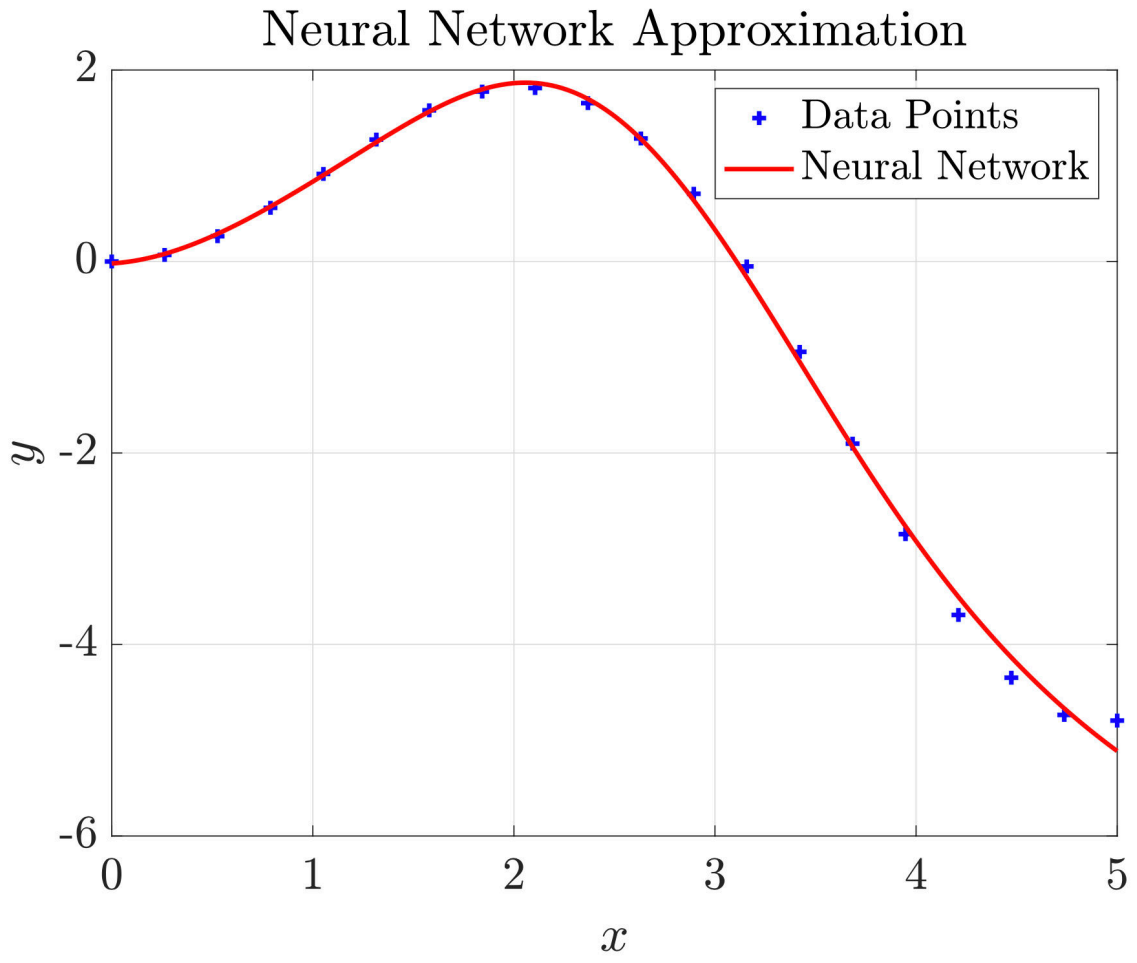
Westny, T., Frisk, E., Olofsson, B: "Vehicle Behavior Prediction and Generalization Using Imbalanced Learning Techniques", IEEE International Intelligent Transportation Systems Conference, 2021

Example: Simple Neural Network for Regression

- Given 20 samples from the **nonlinear function**

$$y = x\sin(x),\ \ x \in [0, 5]$$

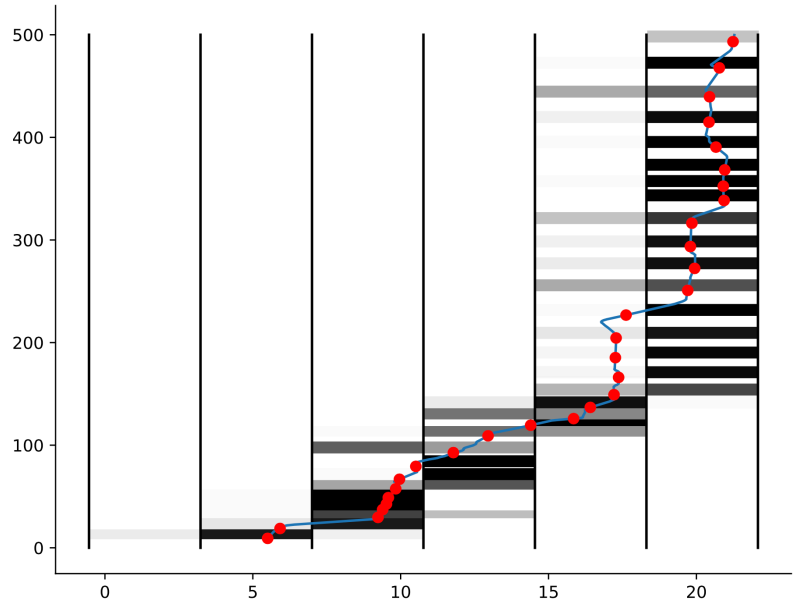
- A **neural network approximation** with **one hidden layer** with 30 neurons shown in red.



Hand-in Exercise 5: Neural Network for Intent Prediction

- In Hand-in Exercise 5, **lane-change predictions** are computed based on **driver data** from the I-80 highway section in the U.S.
- A **neural network** is trained as a **classifier** using 41 features in 4 383 trajectories.

$$y = \begin{cases} 0 & \text{if the vehicle will change to the left within three seconds} \\ 1 & \text{if the vehicle will stay in lane for the next three seconds} \\ 2 & \text{if the vehicle will change to the right within three seconds} \end{cases}$$



I-80 data set citation: U.S. Department of Transportation Federal Highway Administration. (2016). Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data. [Dataset]. Provided by ITS DataHub through Data.transportation.gov. Accessed 2020-09-29 from http://doi.org/10.21949/1504477.

Introduction to Reinforcement Learning

Introduction to Reinforcement Learning (1/2)

- In some environments or scenarios, it is **difficult to explicitly model the dynamics** (e.g., an autonomous vehicle in an unstructured environment).
- How to find **control laws** under **uncertainty** and (partially) **unknown dynamics,** and in addition possibly **uncertain state information**?
- In **control engineering**: system identification, control (e.g., adaptive and stochastic), and observer design.
- **Reinforcement learning** has successfully been used for various challenging scenarios (computer games, chess, Go, robot control, etc.).
  - Often scenarios with a **closed world** and, e.g., **distinct rules** like in games.

Introduction to Reinforcement Learning (2/2)

- **Learn** how to act optimally in each state in the world model (a policy), by interacting with the environment and **maximize an accumulated received reward signal**.
- Learn which actions to take in different situations by **sequentially taking actions and exploring the environment**.
- Trade-off between ***exploration*** and ***exploitation***.
- **Uncertain environments**, typically modeled using probability distributions.

A First Example – K-Armed Bandit (1/4)

- A classical first example in reinforcement learning is the so called **k-armed bandit problem**.
- Consider choosing between ***k* = 3 different actions** at every time step (corresponding to 3 different arms on a slot machine).
  - Each **action results in a reward** according to an ***a priori* unknown normal distribution** with constant mean and variance.
- The **objective** is to **maximize the accumulated reward** over a specified number of time steps.

A First Example – K-Armed Bandit (2/4)

- Let us try to decide on the **optimal action (policy)** by interacting with the 3-armed bandit, i.e., **sequentially choosing and applying actions and observing the rewards obtained**.
- Introduce a function that **measures the value of a particular action**:

$$Q_t(a) = \frac{\text{sum of rewards when action } a \text{ chosen}}{\text{total number of times action } a \text{ chosen}}$$

- Actions chosen according to an **epsilon-greedy policy**, where a **random action** is chosen with probability ε (exploration) and else the currently **greedy action** (exploitation) is chosen as

$$a_{\text{greedy}} = \arg\max_{a'} Q_t(a')$$

A First Example – K-Armed Bandit (3/4)

Iteration	$Q_t(a_1)$	$Q_t(a_2)$	$Q_t(a_3)$	$R_t(a_1)$	$R_t(a_2)$	$R_t(a_3)$	a	Type
1	4.8510	6.1812	5.2415	4.8510	6.1812	5.2415	1,2,3	*Initial*
2	4.8892	6.1812	5.2415	4.9273			1	*Explore*
3	4.8892	5.6798	5.2415		5.1784		2	*Exploit*
4	4.8892	5.6487	5.2415		5.5864		2	*Exploit*
5	4.8892	5.6487	4.8661			4.490	3	*Explore*
6	4.8892	5.4258	4.8661		4.7572		2	*Exploit*
7	4.8892	4.9476	4.8661		3.0346		2	*Exploit*
8	4.8892	4.9476	5.9204			8.029	3	*Explore*
9	4.8892	4.9476	6.0912			6.603	3	*Exploit*
10	4.8892	4.9476	6.4277			7.773	3	*Exploit*
11	4.9536	4.9476	6.4277	5.0826			1	*Explore*
12	4.9536	4.9476	6.1285			4.632	3	*Exploit*

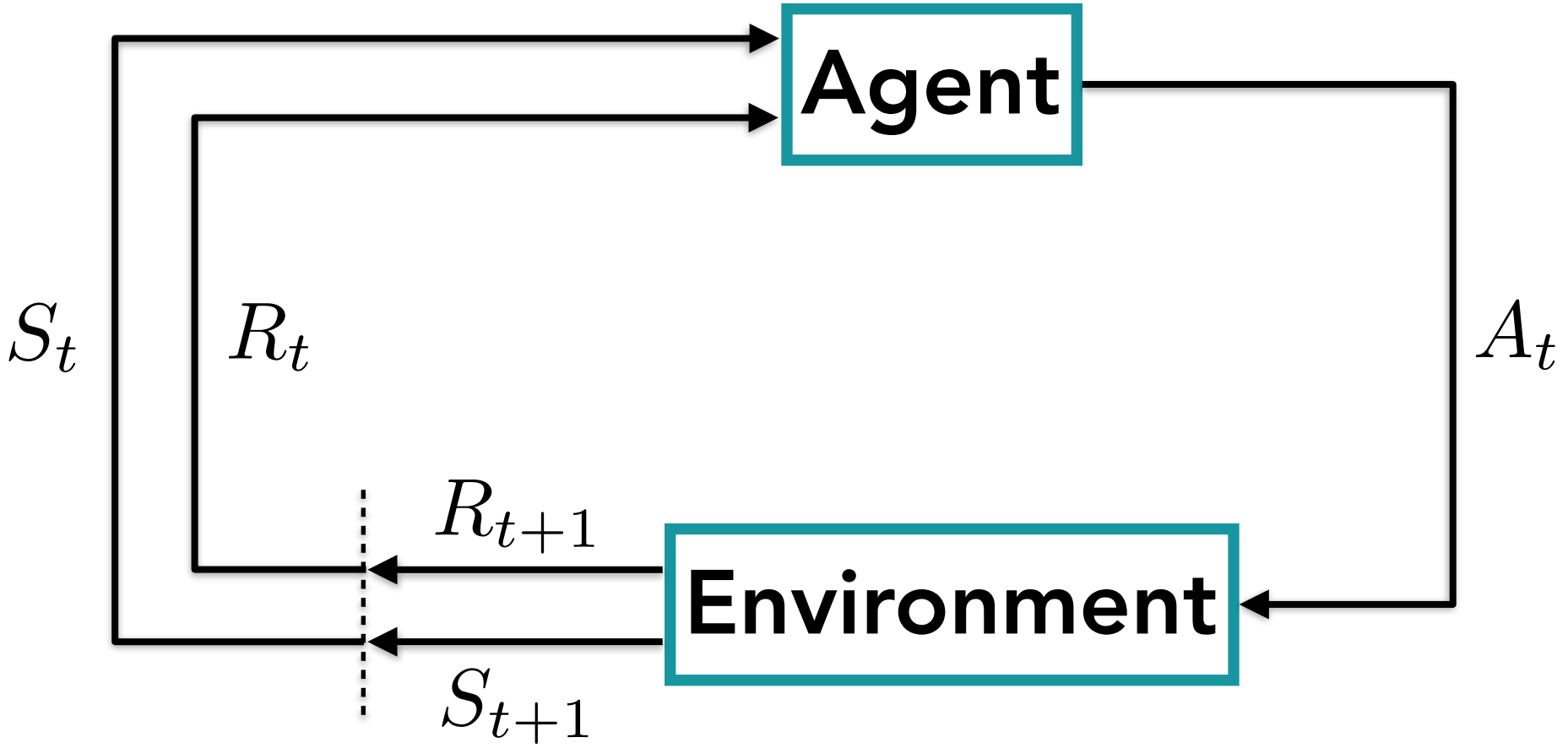
A First Example – K-Armed Bandit (4/4)

- The **actual parameters** of the normal distributions of the 3-armed bandit example are (*mean and standard deviation*):

$$\text{Action 1: } \mu_1 = 5.5,\ \sigma_1 = 1 \quad \text{Action 2: } \mu_2 = 5,\ \sigma_2 = 1 \quad \text{Action 3: } \mu_3 = 6,\ \sigma_3 = 1$$

- The example introduces many **key features of reinforcement learning**:
  - **Exploration/exploitation** to both gain and use information, estimate **value of an action** by repeated interactions by the agent with the environment, and **stochastic uncertainty** in the world.
- In this example, the **reward is not dependent on a state**. This will be considered in the formal definition of a ***Markov decision process***.

The Agent-Environment Interaction Model



- **Searching** for a **policy** (control law) on how to act in each state.

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Definition of Concepts

- **Agent** – the controller subject to learning.
- **Environment** – the agent interacts with the surroundings (controlled system).
- **Action** $A_t \in \mathcal{A}(s)$ – control signal decided by the agent.
- **State** $S_t \in \mathcal{S}$ – describes all relevant aspects of the environment.
- **Reward** $R_t \in \mathcal{R}$ – numerical value given by the environment and received by the agent as a result of the action taken.

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Markov Decision Process (MDP) (1/3)

- **Markov decision processes** (MDPs) are a mathematical framework for sequential decision-making problems.
- Basis for formulation of many reinforcement learning problems, here we consider **finite MDPs**.
- Maximize the (discounted) **accumulated return**

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Markov Decision Process (MDP) (2/3)

- **Policy** defines **action** to be taken (updated sequentially)

$$a = \pi(s), \quad \text{or} \quad \pi(a|s) = \mathrm{P}(A_t = a|S_t = s)$$

- The **state-value function** defines expected return, given policy

$$v_\pi(s) = \mathrm{E}_\pi(G_t|S_t = s) = \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s\right)$$

- The **action-value function** defines the expected value of an action in a certain state

$$q_\pi(s, a) = \mathrm{E}_\pi(G_t|S_t = s, A_t = a) = \mathrm{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a\right)$$

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Markov Decision Process (MDP) (3/3)

- The **state transition** resulting from an action described by

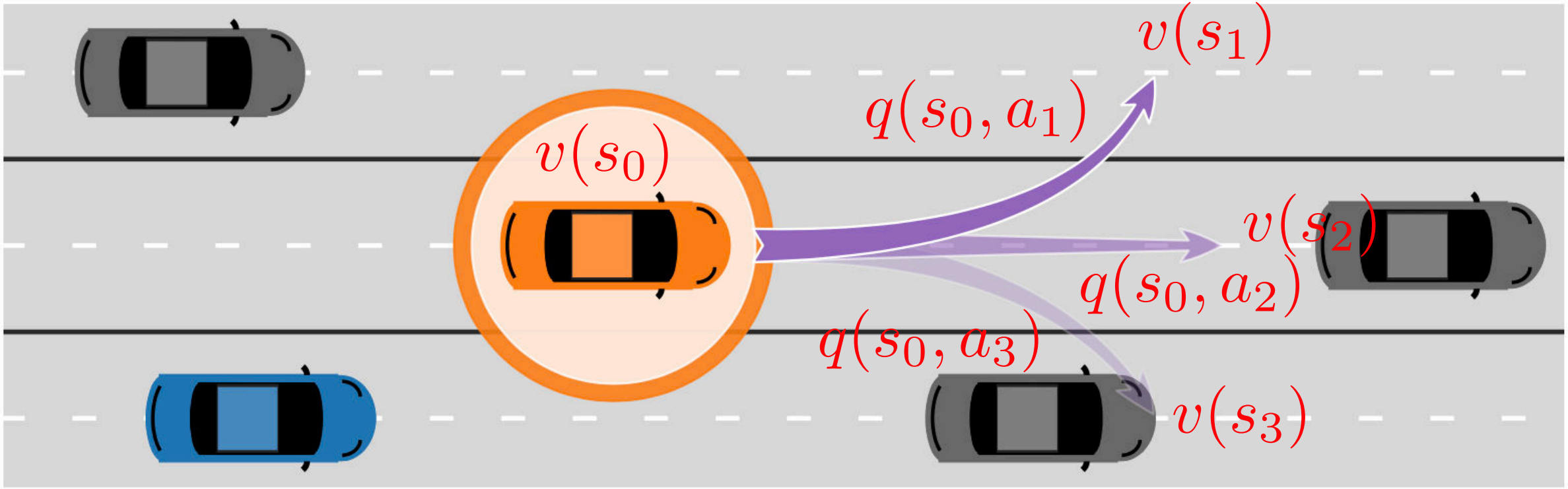
$$p(s', r|s, a) = \mathrm{P}(S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a)$$

- The MDP consists of the following (finite) **sets**, **reward signal**, and **state-transition and reward probabilities**

$$\mathcal{S},\ \mathcal{A},\ \mathcal{R},\ p(s', r|s, a)$$

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

State and Action-Value Functions – Example



The Markov Assumption

- The **Markovian assumption** implies that all information needed is contained in the **current state**.
- In terms of **probabilities**, this can be expressed as

$$\mathrm{P}(S_t, R_t | A_{t-1}, S_{t-1}, \ldots, A_0, S_0) = \mathrm{P}(S_t, R_t | A_{t-1}, S_{t-1})$$

- This assumption is **fundamental** in the formulation of the MDP, since state and action-value functions depend only on **the current state**.

Partially Observable Markov Decision Process (POMDP)

- In the MDP formulation, it has so far been assumed that the current state is **fully observable**.
- What if all states are not known, or cannot be measured?
- **Partially observable MDP** (**POMDP**) takes the uncertainty in the state into account, by introduction of a **belief state** that is updated based on observations.

Åström, K. J.: "Optimal control of Markov processes with incomplete state information", Journal of Mathematical Analysis and Applications, 10(1), 174-205, 1965.

Optimality

- In reinforcement learning, **policies maximizing the total reward** are searched for

$$\pi_* = \text{argmax}_\pi v_\pi(s)$$

- The **optimal value function** and **optimal action-value function** are given by

$$v_*(s) = \max_\pi v_\pi(s), \quad q_*(s,a) = \max_\pi q_\pi(s,a)$$

- Must hold all for **all states and all allowed state-action pairs**.

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

The Bellman Optimality Equations

- The **Bellman optimality equations** define recursive relationships for value function and action-value functions.
- Optimality equation for the **value function** ($s \to s'$)

$$v_*(s) = \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) = \max_a \sum_{s',r} p(s', r|s, a)(r + \gamma v_*(s'))$$

- Optimality equation for the **action-value function** ($s \to s'$)

$$q_*(s,a) = \mathrm{E}\left(R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') | S_t = s, A_t = a\right)$$

$$= \sum_{s',r} p(s', r|s, a)(r + \gamma \max_{a'} q_*(s', a'))$$

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Dynamic Programming in Reinforcement Learning

- **Dynamic programming** (DP) can be used to solve a problem formulated as a finite MDP, given **exact model knowledge**.
- Basic idea: **Search for optimal policies** using the **Bellman optimality equations** for the value or action-value functions.
- Two main variants: **policy iteration** (initialize, policy evaluation, policy improvement, repeat) and **value iteration** (next slide).

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Value Iteration towards an Optimal Policy

Algorithm 1: Value Iteration

Initialize $V(s)$ arbitrarily $\forall s \in \mathcal{S}$
repeat
  $\Delta = 0$
  foreach $s \in \mathcal{S}$:
    $v = V(s)$
    $V(s) = \max_a \sum_{s',r} p(s', r|s, a)(r + \gamma V(s'))$
    $\Delta = \max(\Delta, |v - V(s)|)$
until $\Delta < \varepsilon$
return policy π as:
  $\pi(s) = \text{argmax}_a \sum_{s',r} p(s', r|s, a)(r + \gamma V(s'))$

Bellman optimality equation

- **After** convergence, it holds that $V(s) = v_*(s)$.

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Unknown Environment and Exploration vs. Exploitation (1/2)

- Policy and value iterations rely on **known state-transition and reward probabilities**.
- What if these are not known *a priori* and the **environment is unknown**?
- **Learn** the characteristics of the **environment** by **interacting** with it.

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Unknown Environment and Exploration vs. Exploitation (2/2)

- Leads to the question of **trade-off** between **exploration** (test **new actions** to investigate the environment) and **exploitation** (use the **information acquired** so far and act according to **best possible strategy**).
- **Epsilon-greedy exploration** common choice:

$$\pi(a|s) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}|}, & a = \text{argmax}_a Q(s,a) \\ \frac{\varepsilon}{|\mathcal{A}|}, & a \neq \text{argmax}_a Q(s,a) \end{cases}$$

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Temporal Difference (TD) Learning

- **Temporal-difference methods** try to learn optimal policies **without explicit knowledge** of the environment and its dynamics.
- Concepts from **dynamic programming** and **Monte Carlo methods** used and combined, by **bootstrapping** (update estimates based on other learned estimates).
- **Temporal-difference (TD) error**: $\Delta_{\mathrm{TD}} = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$
- Basic idea: **successively update** the **value function** as (α is a parameter)

$$V(S_t) = V(S_t) + \alpha \Delta_{\mathrm{TD}}$$

- Common methods of this type are **SARSA** and **Q-learning** (next slide).

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Q-Learning

$$Q(s,a) \approx q_*(s,a)$$

Algorithm 2: Q-Learning

Initialize $Q(s,a)$ arbitrarily $\forall s \in \mathcal{S},\ a \in \mathcal{A}(s)$
    and **initialize** $Q(S_T,\cdot) = 0\ \forall$ terminal states
repeat for K episodes
    Initialize S to start state
    repeat
        Choose action A from state S using policy determined from Q
        Take action A, receive reward R, and get next state S'
        $Q(S,A) = Q(S,A) + \alpha(R + \gamma \max_a Q(S',a) - Q(S,A))$
        $S = S'$
    until S is a terminal state

Temporal-difference (TD) error

- **Q-learning** is called an off-policy TD method; **SARSA** is a common on-policy method.

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Value-Function and Action-Value Approximations

- When the state and action spaces **increase in dimension**, **parameterized functions** can be used to **approximate** the value function or action-value function

$$v_*(s) \approx \hat{v}(s;\theta), \quad q_*(s,a) \approx \hat{q}(s,a;\theta), \quad \theta \text{ parameters}$$

- **Common choices** of functions are linear, polynomials, neural networks, etc.
- Reformulate previous methods to **update parameters** instead of value or action-value functions directly (e.g., using gradient descent).

Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Policy-Gradient Methods

- An alternative method to approximating the value or action-value functions is to **directly parameterize the policy** (possibly along with value function) as

$$\pi(a|s,\theta) = \mathrm{P}(A_t = a|S_t = s, \theta_t = \theta)$$

- Then update parameters θ based on some performance metric $J(\theta)$ (compare with cost function from Lecture 5).
- Common methods in this category are **REINFORCE** and **Actor-Critic**.

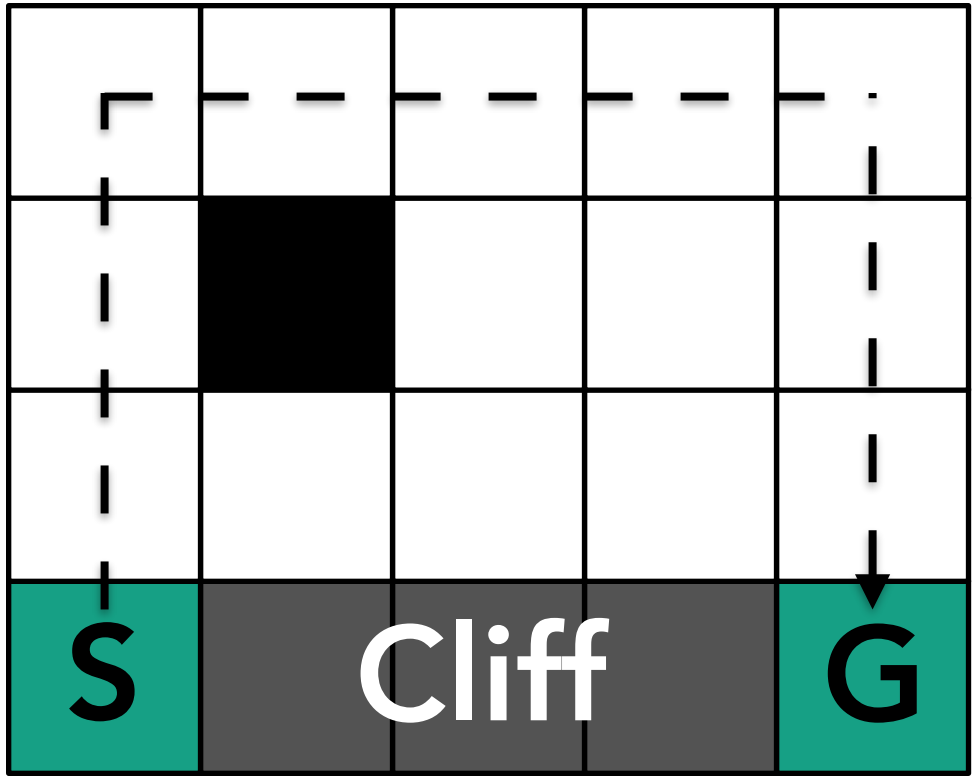
Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Simulation as a Tool for Reinforcement Learning

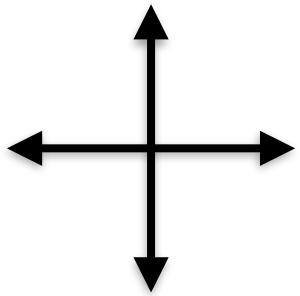
- In some environments, it could be **challenging to iteratively interact** with the environment by **actual experiments**.
- A **simulated environment** can therefore be used to train the algorithm (possibly also **reducing the time** to perform the required experiments).

Reinforcement Learning in Hand-in Exercise 5

- Autonomous vehicle moves in a **grid world** from start to goal.
- **Stochastic winds** might lead to detour from intended action.
- **Cliffs** should be avoided (negative reward).
- **Small negative reward** for states other than those in the lower row.



High negative reward



Possible actions

Expanded version of Example 6.6 in Sutton, R. S., & A. G. Barto: Reinforcement learning: An introduction. MIT Press, 2018.

Combining Neural Networks and Reinforcement Learning – Deep Q-Learning

Introduction

- For cases where it is **infeasible** to use a **tabular Q-function**, approximations using neural networks could be used.
  - Consider, e.g., **continuous states and actions**, discretization often results in very large tabular Q-functions.
- One example of such an approach is **Deep Q-learning**. Introduce the approximation

$$q^*(s, a) \approx Q(s, a; \theta)$$

- The parameter vector θ defines the **neural-network model**, could be **high-dimensional**, but still significantly less than a tabular Q-function.

Mnih, V., Kavukcuoglu, K., Silver, D. et al: "Human-level control through deep reinforcement learning", Nature 518, 529–533, 2015.

Deep Q-Learning (1/3)

- Deep Q-learning builds on **experience replay**, let the agent interact with the environment through several **episodes** (e.g., using an **epsilon-greedy strategy**) and observe the results.
- **Record the outcome** of the actions of the agent at each time step as the **tuple**

$$(\text{state}, \text{action}, \text{reward}, \text{next\_state}) = (S_t, A_t, R_t, S_{t+1})$$

- For increased stability of the method, **two neural networks** with the same structure but different parameter vectors θ and θ^- are used.

Mnih, V., Kavukcuoglu, K., Silver, D. et al: "Human-level control through deep reinforcement learning", Nature 518, 529–533, 2015.

Deep Q-Learning (2/3)

- Recall the **temporal-difference (TD) error**. The parameter vector are updated by minimizing the squared difference between the **target** and the **prediction** as

$$\text{minimize}_\theta \sum_{t=1}^{T} \Big(\underbrace{R_t + \gamma \max_{a'} \hat{Q}(S_{t+1}, a'; \theta^-)}_{\textbf{Target}} - \underbrace{Q(S_t, A_t; \theta)}_{\textbf{Prediction}} \Big)^2$$

- **Gradient descent** used to iteratively updating the parameter vector θ (typically performed sequentially).

Mnih, V., Kavukcuoglu, K., Silver, D. et al: "Human-level control through deep reinforcement learning", Nature 518, 529–533, 2015.

Deep Q-Learning (3/3)

- After a **pre-defined number of steps**, the parameters of the **target neural network** are updated to match the **predictor network**

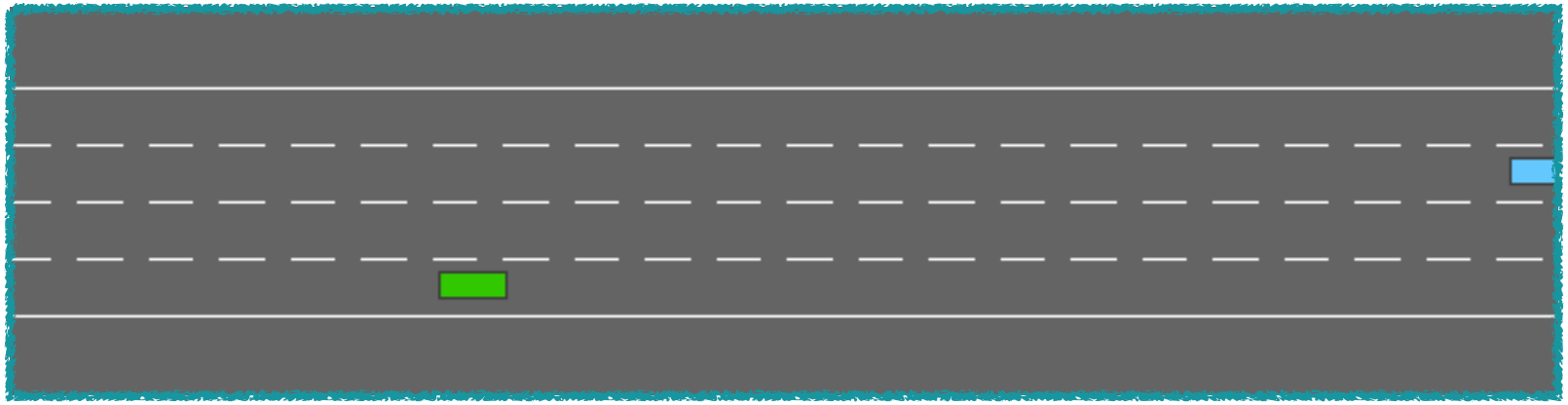
$$\theta^{-} = \theta$$

- **Double deep Q-learning** separates the target and predictor parts even further (not further discussed here).
- **Error clipping** between -1 and 1 further improves the method.

Mnih, V., Kavukcuoglu, K., Silver, D. et al: "Human-level control through deep reinforcement learning", Nature 518, 529–533, 2015.

Example: Deep Q-Learning in Highway Scenario (1/3)

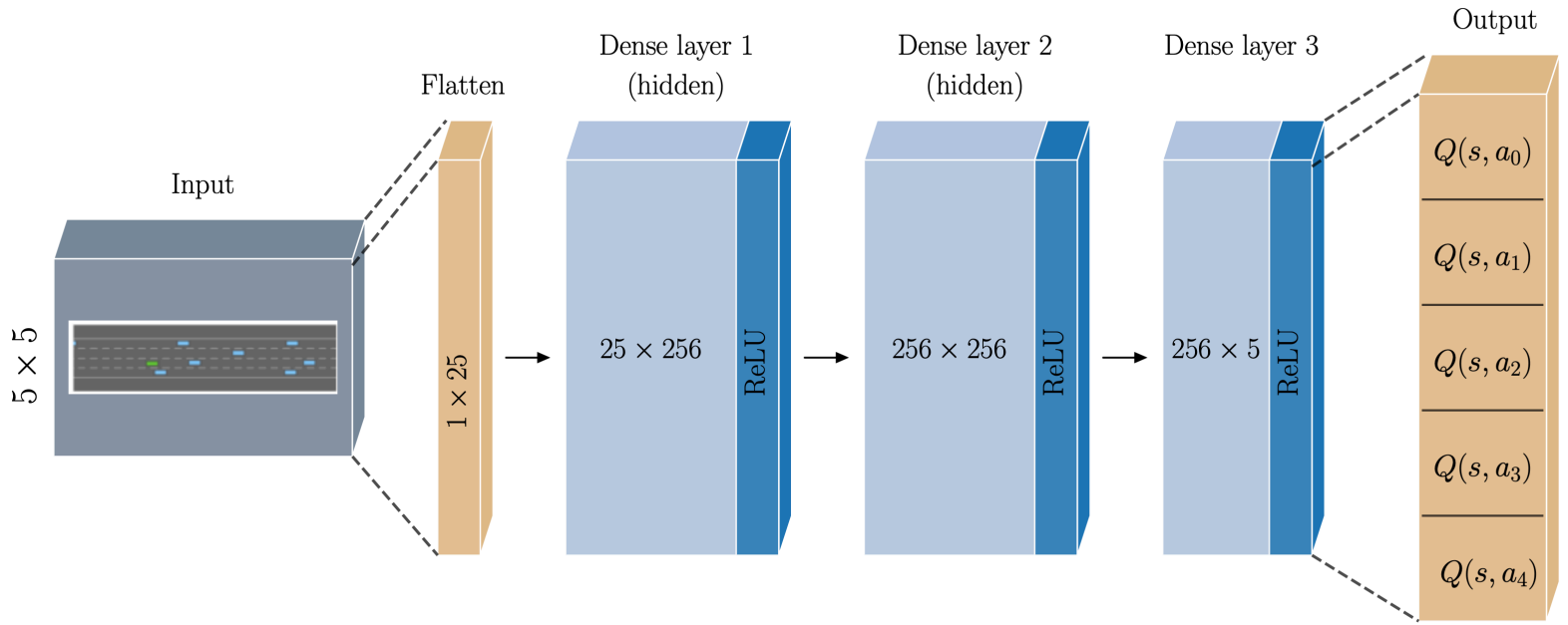
- **Highway scenario** used in extra assignment in **Hand-in Exercise 5**.



- **Navigate** through the lanes, with **multiple vehicles** driving, states and actions are **continuous variables**.
- **Even with a coarse discretization** of the continuous variables, Q-learning with a discrete, tabular Q-function would be **infeasible**.

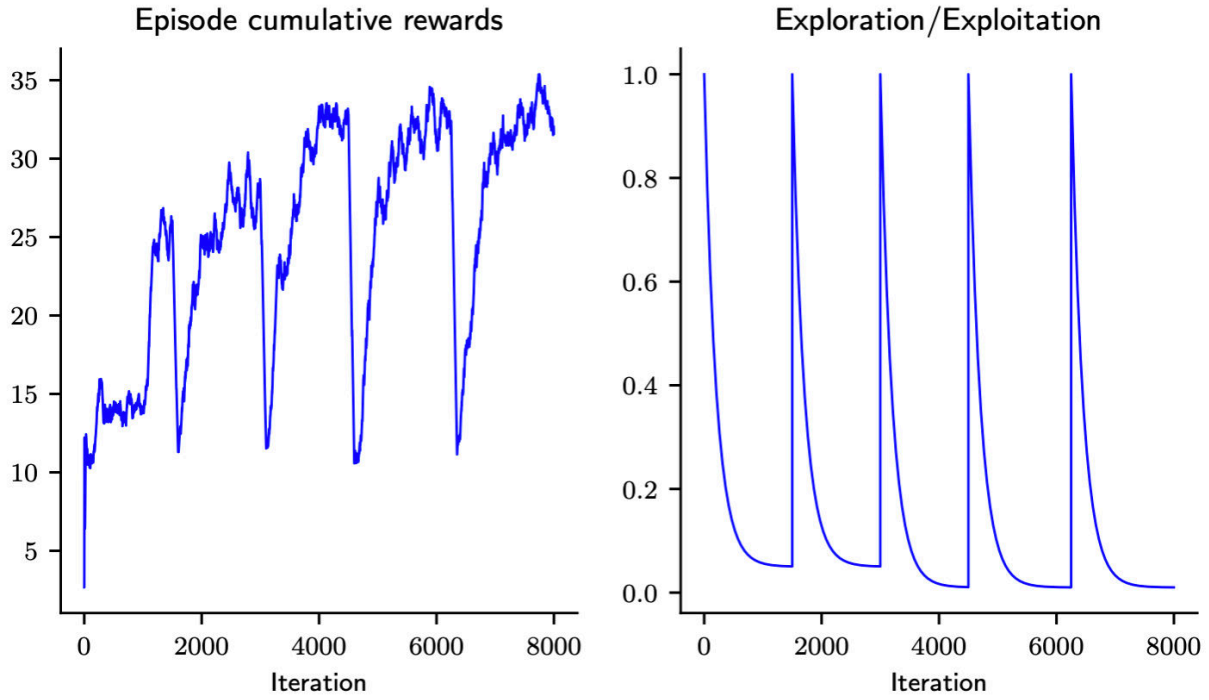
Example: Deep Q-Learning in Highway Scenario (2/3)

- Instead apply **deep Q-learning** using a **neural-network approximation**.
- **Neural network** has **73 753 parameters**. Quite a lot, but still significantly less than the size of a corresponding discretized table.



Example: Deep Q-Learning in Highway Scenario (3/3)

- The agent chooses actions according to an **epsilon-greedy strategy** during training, gives a trade-off between **exploration** and **exploitation**.
- **Sequences** of such phases during training.



Software Libraries for Machine Learning

Some Tools for Neural Networks

- **TensorFlow** open-source library for **machine learning**, notably neural networks with high-dimensional parameter vectors and large amount of data.
  - https://www.tensorflow.org, https://playground.tensorflow.org/
  - **Keras** is a high-level interface to TensorFlow.
- **PyTorch** is an open-source library for machine learning with support for **deep neural networks**.
  - http://pytorch.org/

Toolkit for Reinforcement Learning

- **OpenAI Gym toolkit** for **reinforcement learning** (collection of **example problems and environments**, and **interfaces** to other libraries such as TensorFlow).
  - https://gym.openai.com
- **Environments** for implementation and evaluation of **reinforcement-learning methods** for **traffic scenarios**:
  - https://github.com/eleurent/highway-env

References and Further Reading

All the following books and articles are not part of the reading assignments for the course, but cover the topics studied during this lecture in more detail.

- Goodfellow, I., Y. Bengio, & A. Courville: *Deep Learning*. MIT Press, 2016.
- Hastie, T., R. Tibshirani, J. Friedman, & J. Franklin: *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd Edition, Springer, 2005.
- Sutton, R. S., & A. G. Barto: *Reinforcement learning: An introduction*. MIT Press, 2018.
- Mnih, V., Kavukcuoglu, K., Silver, D. et al: "Human-level control through deep reinforcement learning", Nature 518, 529–533, 2015.
- Westny, T., Frisk, E., Olofsson, B: "Vehicle Behavior Prediction and Generalization Using Imbalanced Learning Techniques", IEEE International Intelligent Transportation Systems Conference, 2021.
- Åström, K. J.: "Optimal control of Markov processes with incomplete state information", *Journal of Mathematical Analysis and Applications*, 10(1), 174-205, 1965.

www.liu.se

LINKÖPING UNIVERSITY