Index

action
  ϵ-greedy algorithm, 409
  greedy action, 409
  pure exploration, 409
activation functions, 87
actor-critic, 410
AdaGrad, 310
ADAM, 316
AdaMax, 317
adjoint equations, 429, 431, 432
asynchronous gradient descent, 446
attention, 203
average loss function, 71

backpropagation, 93, 116, 132, 463
  through time, 189
batch normalization, 151, 200
Bellman equation, 399, 402, 407
bidirectional recurrent neural networks, 198
binary cross entropy, 43

channels, 225
  multiple, 228
  single, 226
Chernoff bound, 481
clipping function, 78, 247
computational cost, 132, 443, 462, 464
computational graph, 466
confusion zone, 310
convergence rate stochastic gradient descent, 297
convolutional neural networks (CNN), 213
cross-validation, 164

deep neural network (DNN), 266
deep reinforcement learning (DRL), 406
define-and-run, 120, 465
define-by-run, 120, 464
dense space, 261, 489
distributed training, 443
dropout, 141, 199
dual representation, 61
dual variables, 61, 65

Elman networks, 184
encoder-decoder, 209, 236
epoch, 296
evidence lower bound (ELBO), 234
exponential moving average, 313

feature importance, 169
feature permutation, 171
feature space, 62
feed forward networks, 69
forward-mode differentiation, 463

gated recurrent units, 195
generalization, 12, 13, 17, 141, 351, 384, 387
generative adversarial network (GAN), 239
gradient descent (GN), 108, 109
Gram matrix, 63
graphical processing unit (GPU), 126, 444, 445, 465

head, 203

high-performance computing (HPC), 450
hinge loss, 60
hyperbolic tangent function, 87

inductive bias, 9
inequalities, 487
initialization
  He/Kaiming, 326
  Xavier, 322

Jordan networks, 183

kernel, 62
kernel perceptron, 63

layer normalization, 201
linear regression, 27, 309
logistic function, 87
logistic loss, 60
logistic regression, 39
long-short-term memory (LSTM), 196

mask, 148
masked self-attention, 209
mean field regime, 355
Mercer theorem, 64
message passing interface (MPI), 449
minibatch, 112, 157
mode collapse, 245
momentum method and SGD, 308
multi-layer neural network, 129, 147
multihead self-attention, 204

Nesterov method, 289
neural ODEs, 427
neural SDEs, 433
neural tangent kernel (NTK), 326
Newton method, 275

objective function, 106
one-hot encoding, 55, 107, 110, 131

padding, 229
parallel efficiency, 448
Pearson correlation, 217
perceptron, 60
perfectly parallel, 446, 448
permutation equivariant, 206
point-to-point communication, 453
Polyak method, 286
polynomial regression, 165
position encoding, 209
prompt, 203
PyTorch, 120

Q function, 399, 403
Q-learning, 399, 406, 408

recurrent neural networks (RNNs), 181
regularization, 137
reinforcement learning, 393
ReLU, 69, 88
reverse-mode differentiation, 462
ridge regression, 138, 141
Riesz representation theorem, 488
RMSProp, 311

self-attention, 203
Shapley value, 173
shattering, 389
Skorokhod space, 479
softplus, 90
steepest descent, 107
stochastic gradient descent (SGD), 105, 129, 293
stochastic process, 475
stochastically bounded, 335
Stone-Weierstrass theorem, 490
stride, 223, 230
strong scaling, 448
synchronous gradient descent, 445

Tauberian theorem, 314
TensorFlow, 120
tightness, 477
time series, 181
token, 203
training, 160
transformer, 207
truncated backpropagation through time (tBPTT), 191
truth tables, 72

universal approximation theorems, 259, 266

validation, 161
vanishing gradient problem, 102, 133
variational auto-encoder, 235, 238
variational inference, 234
VC dimension, 389

Wasserstein GAN, 246
weak scaling, 448
weight initialization, 322

zero-one loss, 60