REINFORCEMENT LEARNING AND STOCHASTIC OPTIMIZATION
A unified framework for sequential decisions

Warren B. Powell

August 22, 2021
Active learning, 48
Actor-critic algorithm, 748
Ad-clicks, 375
Adaptive learning algorithms, 166
as sequential decision, 167
designing policies, 172
learning problems, 167
Affine function, 613
Aggregation, modeling, 102
Aleatoric uncertainty, 463
Algorithm
approximate value iteration
linear model, 732
Benders for two-stage, 786
indifference zone selection, 313
optimal computing budget allocation, 314
temporal-difference learning
infinite horizon, 688
actor-critic, 749
ADP with pre-decision state, 721
approximate policy iteration, 745, 749
linear model, 745
LSTD, 748
backward ADP
lookup tables, 660
parametric model, 661
backward dynamic programming, 615
bias-adjusted Kalman filter
stepsize, 242
direct lookahead policy simulator, 824
double-pass ADP, 726
Gauss-Seidel variation, 622
hybrid value/policy iteration, 627
least squares temporal differencing, 748
MCTS algorithm
backup step, 840
roll-out policy, 838
simulation step, 840
tree policy, 839
policy evaluation, 681
policy iteration, 626
Q-learning
finite horizon, 723, 724
relative value iteration, 622
shortest path algorithm, 60
value iteration, 621
American option, 739
Anytime problems, 33
Apparent convergence, 245
Applications
ad-clicks, 261, 302, 375
aid in Africa, 670
American option, 739
American options, 44
asset selling, 43, 542
asset valuation, 358
bidding, 822
blood management, 368, 587, 767, 888
blood sugar, 543
budgeting problem, 618
building management, 885
business, 3
carbon monoxide, 466
cash management, 154
chemical diffusion, 468
clinical trials, 44, 668
computer simulations, 261
cost-minimizing newsvendor, 185
COVID pandemic, 472
currency exchange, 420
demand estimation, 466, 468
diabetes medication, 401
diabetes medications, 307
diseases, 94
drones, 476, 501
dynamic assignment, 365
dynamic resource allocation, 767
dynamic shortest path, 580
e-commerce, 3, 186
economics, 3
electric utility, 771
electricity contracts, 417
electricity price forecasting, 464
electricity prices, 483
energy, 767
energy forecasting, 401
energy investments, 153
energy modeling
crossing times, 486
jump-diffusion, 484
mean reversion, 483
quantile distributions, 485
regime shifting, 486
energy planning, 817
energy storage, 387, 428, 501, 543, 668
active learning, 430
passive learning, 430
rolling forecast, 588
rolling forecasts, 431
time series price, 429
energy storage policies, 514
CFA, 515
comparison, 516
DLA, 515
hybrid, 516
PFA, 515
VFA, 515
engineering, 3
engineering design, 153
European options, 44
field experiments, 262
finance, 3
financial trading, 501, 582, 817
fleet management, 62
multiple drivers, 63
nomadic trucker, 62
flu mitigation, 867
controllable truth, 870
designing policies, 875
resource constrained, 870
spatial model, 871
static model, 867
time-varying, 869
two-agent learning, 871
with drift, 869
freight transportation, 3
gambling problem, 616
games, 421
graph problems, 355
grid model, 468
homeland security, 44
HVAC controller, 885
model, 886
policies, 887
inventory, 3, 9, 57, 186, 360, 502
investment banking, 771
laboratory experimentation, 374
load matching, 578
logistics, 659
machine replacement, 44
machine scheduling, 822
manufacturing, 268, 772
materials science, 3, 261, 307, 817
medical, 421
medical decisions, 64, 94, 262, 332, 373, 476, 817
medical diagnostics, 466
medical research, 3
money, 767
nested news vendor, 185
news vendor, 55, 156, 331, 770
nomadic trucker, 102, 356
oil inventories, 476
online bookseller, 772
police patrolling, 501
power grid, 466, 470
power transformers, 417
pricing, 63, 93, 469
public health, 3, 94
rat diet, 469
resource allocation, 365
multi period, 773
two-stage, 770
ride hailing, 822, 835
rockets, 38
scientific exploration, 65
shortest path, 60, 156, 822, 830
deterministic, 60
dynamic, 61
robust, 61
stochastic, 60
shuttle bus, 543
SpaceX, 38
sports, 94, 261
stochastic shortest path, 355
stopping problems, 43
supply chain, 3
tic-tac-toe, 742
transformer replacement, 357
transportation, 261, 659
trucking, 401, 659, 767, 822
two-agent news vendor, 881
water reservoir, 502
Approximate dynamic programming, 39
single state, 705
LSPE, 707
LSTD, 706
recursive least squares, 707
Approximate linear programming method
linear model, 753
Approximate policy iteration, 744
linear model, 745
lookup tables, 745
Approximate value iteration, 683
backward pass, 726
bounding $1/n$ convergence, 708
geometric view, 701
gradient-based methods, 697
linear models, 697, 732
lookup tables, 721
post-decision state, 724
pre-decision state, 721
slow backward learning, 686
TD(0), 685
Approximation architectures, 135
Asset valuation, 358
Backgammon, 403
Backpropagation, 726
Backward ADP
benchmarking, 667
aid in Africa, 670
clinical trials, 668
energy storage, 668
finite horizon, 656
lookup table, 660
parametric model, 661
VFA approximation, 664
linear model, 664
monotone function, 665
Backward dynamic programming, 615
Base model, 432
Base models vs. lookahead models, 431
Basis functions, 779
American option, 741
approximate linear programming, 754
tic-tac-toe, 743
Bayesian prior, 92
Bayesian updating
correlated normal beliefs, 93
independent beliefs, 92
normal-normal, 92
precision, 93
Behavior policy, 171, 735
Belief MDP, 880
Belief state, 333, 406
Bellman error, 684
Bellman’s equation, 608
curse of dimensionality, 37
deterministic, 609
expectation form, 610
linear model, 690
linear operator, 613
matrix form, 611
operator form, 613
standard form, 609
Bellman’s optimality equation, 37
Benders decomposition, 785
asymptotic analysis, 787
two-stage, 783
Benders’ decomposition
regularization, 787
Beta distribution, 482
Bias and variance, 97
Bias in max operator, 750
Blood management problem, 368
Boltzmann exploration, 276, 514
Budgeting problem, 618
Chance constrained programming, 49
Classification of problems, 432
state-dependent, cumulative
reward, 434
state-dependent, final reward, 435
state-independent, cumulative
reward, 433
state-independent, final reward, 433
Competitive analysis, 320
Conditional value at risk, 428
Contextual bandit, 332
Control law, 39
Correlated beliefs, 93
Cost function approximation
constraint-modified, 585
blood management, 587
dynamic trading, 582
energy storage, 588
parameterized policy, 592, 832
SPSA, 592
tuning, 593
general form, 577
heat maps, 593
interval estimation, 576
uncertainty bonus, 576
objective-modified, 578
cost function correction, 578
dynamic assignment, 578
pure exploitation, 576
shortest path, 580
Cost-to-go function, 38
COVID pandemic, 472
Crossing times, 486
Cumulative reward, 34
adaptive policy, 327
bandit problems, 264
cumulative loss, 321
examples, 47
expectation
state-dependent, 434
expensive experiments, 286
knowledge gradient, 296, 302, 327
learning, 170
multiarmed bandit, 45
news vendor, 56
policy evaluation, 525, 826
simulated
state-dependent, 436
state-independent, 436
state-independent, 433
stochastic gradient algorithm, 184,
261, 268
stochastic search, 155
two-agent, 872
Curse of dimensionality, 37, 133
Cutting planes, 783
Data driven, 56, 64
Decision tree, 35, 606
Decision variables, 52
constraints, 413
execution decisions, 412
policies, 414
strategic decisions, 412
tactical decisions, 412
types, 411
Decisions, 410
Deep neural networks, 126
Derivative-free stochastic search
CFA policies, 274
Bayes greedy, 274, 276
greedy policy, 274
Interval estimation, 276
interval estimation, 275
UCB policy, 275
designing policies, 322
Direct lookahead
  multiperiod deterministic, 290
  multiperiod stochastic, 292
  restricted multiperiod, 289
  single period lookahead, 287
direct lookahead, 285
exogenous state information, 331
large choice sets, 329
learning in batches, 310
policy function approximation, 273
scaling, 324
tuning, 325
VFA policy, 276
  backward ADP, 279
  Gittins indices, 281
VFA-policy
  Beta-Bernoulli, 278
Designing a policy, 517
Diagnostic uncertainty, 460
Direct lookahead approximations, 505
Direct lookahead model
  approximation strategies, 813
  horizon truncation, 814
  latent variables, 814
  policy approximation, 815
  sampling, 814
  stage aggregation, 814
lookahead objectives, 817
  chance-constrained, 820
  managing risk, 817
  model discounting, 822
  risk measures, 817
  risk-adjusted policies, 819
  robust optimization, 820
  utility functions, 822
notation, 812
Direct lookahead policy
  approximate dynamic
    programming, 811
  approximate lookahead, 812
  backward ADP lookahead policy, 835
  creating lookahead model, 811
  deterministic, 828
discounted, 822
DLA lookahead policy, 835
fixed lookahead policy, 824
lookahead CFA, 834
lookahead PFA, 833
lookahead policy, 810
MCTS, 810, 836
model predictive control, 810
optimal, 808
policy-within-a-policy, 810
robust optimization, 811
rolling horizon procedure, 810
rollout policy, 811
shortest path, 830
stochastic programming, 810
why use, 827
Distributional uncertainty, 481
Double-pass ADP algorithm, 726
Dynamic assignment problem, 365
Dynamic program
  budgeting problem, 618
  gambling problem, 616
Dynamic programming
  model-free, 422, 723
Endogenous learning, 83, 86
Energy storage model, 428
  active learning, 430
  passive learning, 430
  rolling forecasts, 431
  time-series prices, 429
Energy storage policies, 514
  CFA, 515
  DLA, 515
  Hybrid, 516
  PFA, 515
  VFA, 515
Episodic, 681
Epistemic uncertainty, 463
Epsilon-greedy, 514
Evaluating policies, 314
  competitive analysis, 320
  dynamic regret, 319
  empirical performance, 314
  indifference zone, 320
  opportunity cost, 319
  probability of correction selection, 321
  quantiles, 316
static regret, 316, 317
subset selection, 322
excitation, 514
Excitation policy, 553
Exogenous information, 52, 414
  adversarial processes, 420
deterministic, 420
lagged, 417
modeling, 414, 418
scenarios, 416
  supervisory processes, 420
Exogenous learning, 86
Expectation
  compact form, 32
  expanded form, 32
Expected improvement, 287
Experimental uncertainty, 461
Exploration
  Boltzmann exploration, 514
epsilon-greedy, 514
excitation, 514
  Thompson sampling, 514
Exploration vs. exploitation, 513
Exponential smoothing, 90, 224
Final reward, 34, 154, 169, 269
  asymptotic convergence, 321
  bandit problems, 264
cost function approximation, 576
  expectation
    state-dependent, 433, 434
    state-independent, 433
expensive experiments, 286
knowledge gradient, 287, 296, 297
multiarmed bandit, 47
newsvendor, 56
policy search, 547
  simulated
    state-dependent, 437, 731
    state-independent, 436
stochastic gradient algorithm, 193
stochastic search, 184
Finite-horizon approximations
  steady state, 756
Fitted value iteration, 662, 679
Fleet management problem, 62
  fleet, 63
  nomadic trucker, 62
Forecasts
  rolling, 59, 408, 431, 520, 580, 588
Gambling problem, 616
Games
  backgammon, 403
  chess, 421, 744
computer Go, 421, 744
  maze, 41
  video, 126
  video games, 421
Gaussian process regression, 96
Gittins indices, 281
  normally distributed rewards, 283
Hamiltonian, 38
Hierarchical aggregation, 100
  correlations, 138
  estimation, 106
  estimation, 102
Hyperparameter, 188
Hyperparameters, 482, 483
Implementation policy, 171, 276
Inferential uncertainty, 460
Infinite horizon, 619
  policy iteration, 745
temporal-difference learning, 686
Information acquisition, 373
Interval estimation, 275
Inventory problems, 57
  basic, 360
  batch replenishment, 363
  forecasts, 59
general, 361
  lagged, 362
  with lags, 59
  without lags, 58
Inverse optimization, 87
Jump diffusion, 484
k-nearest neighbor, 123
Kalman filter, 239
Kernel regression, 124
Knowledge gradient, 295
  correlated beliefs, 306
cumulative reward, 302
  final reward, 287, 297
  posterior reshaping, 290
  restricted lookahead, 289
sampled belief, 302
Laboratory experimentation, 374
Lagged information, 417
Lagrangian relaxation, 631
Lasso, 110
Latent variable, 407, 462
Learning
  approximation strategies, 87
  Bayesian
    state variable, 93
  bias and variance, 97
  deep neural networks, 126
  endogenous, 83, 86
  exogenous, 83, 86
  exponential smoothing, 90
  frequentist
    state variable, 92
  Gaussian process regression, 96
  inverse optimization, 87
  k-nearest neighbor, 123
  kernel regression, 124
  linear regression, 108
  local polynomial regression, 126
  lookup tables, 87
    Bayesian, 92
    frequentist, 91
  maximum likelihood estimation, 115
  model-based, 41
  model-free, 41
  nonlinear parametric models, 115
  nonparametric definition, 123
  nonparametric models, 88
  nonstationary, 130
    learning process, 132
    transient truth, 131
    truth, 130
  normal equations, 110
  parametric models, 87
    linear, 88
    nonlinear, 88
  sampled belief models, 116
  Sherman-Morrison
    derivation, 137
    supervised, 83
    support vector machines, 127
Learning policy, 171, 276, 735
Learning process, 132
Learning rate, 193
Least squares policy evaluation, 694, 707
Least squares temporal difference, 693, 706, 747, 748
Leveling algorithm, 777
Linear model
  basis functions, 108
  geometric view, 701
  Lasso, 110
  recursive least squares, 111
  sparse additive, 110
Linear models, 107
Linear programming method, 629, 630
Linear quadratic regulation, 38, 630
Linear regression, 108
  Longstaff and Schwartz, 739
Local polynomial regression, 126
Longstaff and Schwartz, 739
  basis functions, 741
Lookahead dynamic programming, 732
Lookahead model, 432
Lookup tables
  aggregation, 99
  Bayesian updating, 92
  state variable, 93
  frequentist updating, 91
  state variable, 92
  hierarchical aggregation, 100, 138
LSPE, 694
  single state, 696
LSTD, 693
  single state, 696
Markov decision process, 36, 605
  average reward, 628
  backward dynamic programming, 615
  Bellman’s equation, 37, 608
  cost-to-go function, 611
  curse of dimensionality, 37
  finite horizon, 613
  infinite horizon, 619
  linear programming method, 630
  objective function, 608
  optimality equation, 608
  partially observable, 400
  random contributions, 612
  state transitions, 612
transition matrix, 612
value function, 611
value iteration
stopping rule, 621
Martingale model of forecast evolution, 408
Maximum likelihood estimation, 115
MCTS
  backpropagation step, 840
  expansion step, 839
  selection step, 838
  simulation step, 839
  steps, 837
Mean reversion, 483
Measure-theoretic view of information, 438
Medical decision making, 64, 373
MMFE, 408
Model
  base model, 432
  lookahead, 432
  nominal, 432
  policies, 495
  true, 432
  universal, 383
Model first, then solve, 2, 5, 8, 21, 355, 385, 388, 414, 495
Model predictive control, 49
Model-free dynamic programming, 422, 723
Modeling frameworks, 54
Modeling time, 392
Monte Carlo, 478
Monte Carlo sampling, 478
Monte Carlo simulation, 460
  inverse cumulative method, 480
Monte Carlo tree search, 836, 837
  optimistic, 842
Monte-Carlo simulation, 477
Multiagent systems
  blood management, 888
  multiple agent, 890
  communication, 862
  active observations, 862
  architecture, 862
  receiving information, 863
  sending information, 863
  signal distortion, 863
  controlling architecture, 866
dimensions, 860
  communication, 860
  coordination, 860
  learning, 860
  the agents, 860
flu mitigation, 867
  controllable truth, 870
  designing policies, 875
  multiarmed bandit, 876
  neural network policy, 876
  resource-constrained, 870
  spatial model, 871
  static model, 867
  time-varying model, 869
  time-varying with drift, 869
HVAC controller, 885
  model, 886
  policies, 887
  modeling, 863
  two-agent learning model, 871
  transition functions, 874
  two-agent newsvendor, 881
types
  controlling, 861
  environment, 861
  learning, 861
  replenishment, 861
Multiarmed bandit, 45
  active learning, 49, 266
  backward ADP, 279
  best hitter, 47
  best medication, 47
  best path, 47
  beta-bernoulli belief, 278
  Boltzmann exploration, 276
  computer science, 263
  contextual, 331, 332
  derivative-free stochastic opt., 266
  dynamic programming, 276
  flu mitigation, 876
  Gittins, 264
  Gittins indices, 281
  normal rewards, 283
  history, 47
  interval estimation, 275
  interval estimation policy, 46
  nonstationary, 326
  objective function, 46
  one-armed, 262
sequential decision problem, 266
state variable, 46
story, 263
Thompson sampling, 276
transient learning, 327
two-armed, 262
uncertainty bonus, 842
upper confidence bounding, 264, 275
variations, 264
Neural network, 87, 88
actor-critic, 749
batch learning, 90
classification, 88, 200
deep, 42, 297, 537
drone, 501
estimation, 199, 540, 876
four layer, 220
games, 518, 744
learning, 274
limitations, 543, 551, 575, 663
medical, 421
parametric, 117
policy, 270, 546, 805
properties, 202
robot, 518, 538, 744
stepsizes, 233
value function, 40, 602, 749, 805, 877
value functions, 499, 512
Newsvendor
final reward, 56
Newsvendor problem, 55
contextual, 57
cumulative reward, 56
final reward, 56
multidimensional, 57
Nomadic trucker, 62, 356, 403
Nonanticipativity constraint, 45, 166
Nonparametric model
definition, 123
Nonparametric models, 122
k-nearest neighbor, 123
kernel regression, 124
local polynomial regression, 126
Nonparametric statistics
tree regression, 128
Nonstationary learning, 130
learning process, 132
martingale truth, 130
transient truth, 131
Normal equations, 110
Objective function, 52
conditional value at risk, 428
cumulative reward, 170
final reward, 169
performance metrics, 424
probability of correct selection, 321
regret, 171
robust optimization, 428
static regret, 317
value at risk, 428
Observational uncertainty, 460
Off-policy learning, 735
Offline learning, 729
On-policy learning, 735
Online learning, 730
Opportunity cost, 319
Optimal computing budget allocation, 314
Optimal control, 37
control law, 39
cost-to-go function, 38
Hamiltonian, 38
Lagrangian relaxation, 631
linear decision rule, 631
linear quadratic regulation, 38, 630
state equation, 631
stochastic, 39
transition function, 37
Optimal stopping, 42
Optimality equations
proof, 632
Optimistic policy iteration, 745
Optimum-deviation policy, 552
Parametric models
linear, 107
nonlinear, 115
Partial policy evaluation, 745
Partially observable Markov decision
process, 463, 879
Partially observable states, 406
Performance elbow, 826
Policy, 495
affine, 502
behavior, 171, 547, 735
Boltzmann, 514
choosing policy class, 522
complexity tradeoffs, 521
cost function approximation, 503
definition, 496
designing, 517
direct lookahead, 505, 808
designing, 505, 808
energy storage problem, 514
CFA, 515
DFA, 515
Hybrid, 516
PFA, 515
VFA, 515
epsilon-greedy, 514
evaluation, 435, 525
examples, 496
excitation, 514
four classes, 172
four classes of policies, 498
hybrid, 509
implementation, 171, 276, 527, 547
learning, 171, 276, 527, 547, 735
linear decision rule, 502
measurability, 440
optimistic, 745
partial policy evaluation, 745
performance-based, 547
policy function approximation, 501
randomized, 513
roll-out heuristic, 836
sampling the shoulders, 551
soft issues, 527
supervised, 547
target, 171, 547, 735
Thompson sampling, 514
tuning, 526
value function approximation, 504
Policy evaluation, 525
infinite horizon, 681
infinite horizon projection, 683
partial simulations, 682
recurrent visits, 681
temporal difference update, 683
Policy function approximation, 540
affine, 541
affine policy, 541
Boltzmann, 541
contextual, 545
limitations, 518
locally linear, 544
monotone policies, 542
nonlinear, 543
Policy gradient theorem, 560
computation, 560
Policy iteration, 626
hybrid value/policy iteration, 627
infinite horizon, 745
Policy search
correlated knowledge gradient, 555
derivative-free methods, 550
exact derivatives
continuous, 555
discrete, 558
policy gradient theorem, 560
excitation, 553
interval estimation, 554
knowledge gradient, 555
numerical derivatives, 548
optimum-deviation, 552
sampled belief, 554
supervised learning, 562
Policy-within-a-policy, 508
POMDP, 400, 463, 879
belief MDP, 880
Post-decision state, 62, 402
Pricing, 63
Prognostic uncertainty, 460
Q-learning, 723
algorithm, 723, 724
augmented state, 404
deep neural network, 42
expected rewards, 563
linear model, 732
lookup tables, 722
off-policy, 697
reinforcement least squares, 733
temporal-difference updating, 41, 603, 697
updating, 41, 227, 505, 603
Quantile distribution, 485
Random number seed, 478
Randomized policies, 513, 641
Receding horizon procedure, 50
Recursive least squares, 111
derivation, 136
multiple observations, 115
nonstationary data, 113
stationary data, 112
Regret, 171
Reinforcement learning, 40
Q-learning, 41
Resource allocation, 365, 769
blood management, 368
dynamic assignment, 365
exogenous information state, 792
general resource allocation, 773
information acquisition, 373
newsflash, 770
two-stage resource allocation, 770
Response surface method, 274
Risk
robust optimization, 428
Risk measures, 427
conditional value at risk, 428
value at risk, 428
Risk operator, 818
Robust optimization, 50, 428
as a policy, 821
optimization problem, 821
robust policy, 51
uncertainty set, 50
box constraints, 821
Roll-out heuristic, 836
Rolling forecast, 520
Rolling forecasts, 408
Rolling horizon procedure, 50
Sample average approximation, 160
convergence rate, 162
Sampled belief models, 116
Sampled models, 159, 488
Sampling the shoulders, 551
SARSA, 723
Scientific exploration, 65
Sequential kriging, 288
Sherman-Morrison formula, 95, 137
Shortest path
information collecting, 375
stochastic, 355
Shortest path problem, 60
dynamic, 61
robust, 61
stochastic, 60
deterministic, 60
Simulation optimization, 48, 312
indifference zone algorithm, 313
OCBA, 48, 313
Simultaneous perturbation stochastic approximation
see SPSA, 197
Soft max, 514
Software, 385
Specification error, 467
SPSA, 197
mini-batch, 197
tuning, 592
State variables, 51
asset pricing, 358
belief, 333
belief state, 398, 406
definition, 394, 395
dynamic assignment problem, 368
factored, 408
flat representation, 408
gambling problem, 616
information state, 398
lagged, 401
latent variable, 407
physical state, 333, 398
post-decision state, 62, 402
pre-decision state, 402
rolling forecasts, 408
shortest path, 404
three types, 397
Stepsize policies
apparent convergence, 245
BAKF proof, 248
bias-adjusted Kalman filter, 242
convergence conditions, 226
deterministic, 225
constant, 228
harmonic, 229
McClain, 229
polynomial learning rate, 229
search-then-converge, 230
guidelines, 246
infinite horizon bounds, 708
introduction, 191
optimal
approx. value iteration, 243
BAKF, 241
nonstationary - I, 239
nonstationary - II, 240
stationary, 237
optimal stepsizes, 236
stochastic, 231
AdaGrad, 235
ADAM, 234
convergence conditions, 232
Kesten’s rule, 233
RMSProp, 236
stochastic gradient adaptive stepsize, 234
Trigg, 234
Stochastic approximation procedure - see stochastic gradient algorithm, 191
Stochastic decomposition, 785, 787
Stochastic gradient algorithm, 185, 191, 689
as a sequential decision problem, 202
finite differences, 195
for neural networks, 199
gradient smoothing, 194
martingale proof, 210
mini-batches, 197
older proof, 207
second order methods, 195
SPSA, 197
Stochastic modeling
crossing times, 486
electricity prices, 483
energy illustration
crossing times, 486
jump diffusion, 484
mean reversion, 483
quantile distribution, 485
regime shifting, 486
quantile distribution, 485
regime shifting, 486
Stochastic optimal control, 39
Stochastic optimization
asymptotic formulation, 34
cumulative reward, 34
final reward, 34
Stochastic programming, 44
Benders decomposition, 783
interstage independence, 784, 793
lookahead policy, 45
nonanticipativity, 45
nonanticipativity constraint, 45
recourse variables, 45
two-stage, 45
Stochastic search, 32
adaptive learning, 152
cumulative reward, 155, 184
derivative-based, 34
derivative-free, 34
deterministic methods, 152, 155
chance-constrained, 157
Markov decision process, 158
newsvendor, 156
optimal control, 157
shortest path, 156
final reward, 154, 184
newsvendor, 151
sampled approximations, 152
sampled models, 159
chance-constrained, 161
convergence, 162
linear program, 160
parametric model, 161
state-dependent, 155
state-independent, 155
Stopping time, 43
Supervised learning, 83, 744
Support vector machines, 127
Target policy, 171, 735
TD(\(\lambda\)), 684
TD(0)
recursive least squares, 695
TD(0) updating
slow backward learning, 686
Temporal difference learning, 684
Temporal-difference learning, 683
infinite horizon, 686
Thompson sampling, 276, 514
Three curses of dimensionality, 37, 614
Tic-tac-toe, 742
Time, 392
Trajectory following, 735
Transformer replacement, 357
Transient learning model, 327
knowledge gradient, 328
Transition function, 52, 421
batch replenishment, 364
Transition matrix, 612
Tree regression, 128
Trucking, 403
Two-stage stochastic programming, 844
  Benders’ decomposition, 783

Uncertainty
  adversarial, 463
  aleatoric, 463
  algorithmic, 470
  algorithmic uncertainty, 461
  coarse-grained, 463
  communication, 470
  communication uncertainty, 461
  control, 469
  control uncertainty, 461
  diagnostic, 460, 465
  distributional, 463
  epistemic, 463
  experimental, 461, 467
  fine-grained, 463
  goal, 471
  goal uncertainty, 461
  implementation, 469
  inferential, 460, 465
  model, 467
  model specification, 467
  model uncertainty, 461
  observational, 460, 461
  political, 471
  political uncertainty, 461
  prognostic, 460, 463
  regulatory, 471
  transitional, 469
  transitional uncertainty, 461

Uncertainty quantification, 460

Universal modeling framework, 7, 51, 383, 384
  base model, 431
  compact formulation, 54
  decision variables, 52, 410
  energy storage illustration, 428
  exogenous information, 414
  expanded formulation, 51
  lookahead model, 431
  objective function, 52, 424
    risk measures, 427
    uncertainty operators, 427
  problem classifications, 432
  software, 385
  state variable, 394
  state variables, 51
  transition function, 52, 421

Upper confidence bounding, 275

Value at risk, 428

Value function approximation, 63
  Bayesian learning, 703
    correlated beliefs, 704
    creating the prior, 705
    parametric models, 704
  cutting planes, 783
  leveling, 777
  regression methods, 779
    separable piecewise linear, 781
    tic-tac-toe, 742

Value function approximations
  linear approximation, 791
  piecewise linear, 776

Value iteration, 621
  algorithmic update, 719
  bound, 624
  bounds, 624
  error bound, 640
  Gauss-Seidel variation, 622
  monotone convergence, 639
  monotonic behavior, 624
  pre-decision state, 721
  proof of convergence, 636
  relative value iteration, 622
  stopping rule, 621