

Index

- accelerated gradient methods, 55–70, 94
 - for composite nonsmooth optimization, 71, 168
 - for constrained optimization, 126
- accumulation point, 34, 201
- active-set method, 114
- adjoint method, 190–192
 - application to neural networks, 191–192
 - forward pass, 190
 - relationship to chain rule, 190
 - reverse pass, 190
- algorithmic differentiation, *see* automatic differentiation
- alternating direction method of multipliers (ADMM), 181–184, 186, 187
- augmented Lagrangian method, 167, 180–182, 186, 187
 - comparison with dual subgradient method, 181
 - software, 187
 - specification of, 181
- automatic differentiation, 103, 192–195
 - checkpointing, 195
 - computation graph, 193
 - reverse mode, 194
 - reverse sweep, 194, 196
- averaging of iterates
 - in dual subgradient method, 180
 - in mirror descent, 48, 50
 - in the stochastic gradient method, 89
 - in subgradient method, 156, 157
- back-propagation, 75, 194
- boundary point of a set, 211
- bounds, 26, 114, 118, 122, 185, 186, 198
- Bregman divergence, 45–47, 50
 - generating function for, 45
- bundle methods, 156, 168
- cardinality of vector, 150
- Cauchy-Schwartz inequality, 31, 149
- chain rule, 188–190
 - efficiency of, 190
 - forward pass, 189
 - reverse pass, 189
- Chebyshev iterative method, 57, 71
- classification, 2, 12, 14, 77–78, 101, 191
- clustering, 3, 6
- co-coercivity property, 25, 52
- complementarity condition, 196
- complexity, 13, 14, 32, 42, 115, 203–204
 - lower bounds, 56, 70–71
 - of gradient methods, 61
 - of second-order methods, 42–44
- composite nonsmooth function, 146–150, 154, 160
 - first-order necessary conditions, 147
 - first-order optimality conditions, 146–148
 - strongly convex, 147
- compressed sensing, 168
- computational differentiation, *see* automatic differentiation
- condition number, 61
- conditional gradient method (Frank-Wolfe), 127–130, 186
 - definition of, 128
- cone, 119, 201
 - polar of, 201
 - of positive semidefinite matrices, 173
- conjugacy, 69

conjugate gradient method, 55, 68–70
 linear, 68–70, 72
 nonlinear, 70–72
 consensus optimization, 182–184
 constrained optimization, 15, 21, 118–129,
 133, 146, 170, 172, 196
 convex, 144–146
 equality constraints, 170–171, 177, 186,
 195, 197
 statement of, 118, 170
 constraint qualification, 145, 175, 177
 convergence rate
 linear, 30, 33, 35, 105, 126
 Q-linear, 201, 203
 R-linear, 61, 201
 sublinear, 33, 68, 82, 105, 124, 128, 203
 convex hull, 156, 206, 209
 convexity
 of function, 21
 modulus of, 22, 30, 34, 38, 50, 85, 88, 90,
 93, 104, 107, 111, 113, 161, 165, 213
 in non-Euclidean norm, 45, 50
 of quadratic function, 31, 55, 58
 of set, 21, 144, 200, 208
 strong, 21–24, 30–32, 34, 45, 47, 88, 93,
 107, 109, 115, 120, 148, 165
 weak, 21, 107, 112
 coordinate descent methods, 39, 100–114
 accelerated, 115
 block, 100, 101, 113–114, 116, 182
 comparison with steepest-descent method,
 109–111
 cyclic, 110–113, 115
 for empirical risk minimization, 101–102
 for graph-structured objective, 102–103
 in machine learning, 101
 parallel implementation, 116
 proximal, 154, 164–167
 random-permutations, 112
 randomized, 37, 101, 105–111, 115, 165
 for regularized optimization, 113
 Daskin's Theorem, 133, 141–142, 151, 179
 data analysis, 1–3, 100
 data assimilation, 188, 190
 deep learning, *see* neural networks
 descent direction, 27–155
 definition of, 27
 Gauss-Southwell, 37, 115
 in line-search methods, 36–38
 randomized, 37

differential equation limits of gradient
 methods, 56–57, 71
 dissipation term, 56
 directed acyclic graph (DAG), 193
 directional derivatives, 40, 137–141, 153
 additivity of, 138
 definition of, 137
 homogeneity of, 138
 at minimizer, 137
 distributed computing, 183, 184
 dual problem, 170, 172, 178, 185
 for linear programming, 206
 dual variable, *see* Lagrange multiplier
 duality, 170, 171
 for linear programming, 200, 205–206
 strong, 178–179, 206, 207
 weak, 155, 172–174, 206
 duality gap, 173
 example of positive gap, 173–174, 187
 effective domain, 134, 136, 139, 143, 146
 eigenvalue decomposition of symmetric
 matrix, 202
 empirical model, 3
 empirical risk minimization (ERM), 78–80,
 95, 101–102
 and finite-sum objective, 79
 entropy function, 46
 epigraph, 21, 134, 135
 epoch, 160
 Euclidean projection, *see* projection operator
 extended-value function, 134, 144
 Farkas Lemma, 205–207
 feasible set, 118
 feature selection, 2, 5
 feature vector, 1, 192
 finite differences, 103
 finite-sum objective, 2, 12, 77, 80, 81, 85–87,
 94, 96, 183, 184, 192
 frame, 83
 Gauss-Seidel method, 100, 110, 111
 Gelfand's formula, 60
 generalizability, 7, 13
 global minimizer, 27
 Gordan's Theorem, 207, 209
 gradient descent method, *see* steepest-descent
 method
 gradient map, 162

- gradient methods with momentum, *see*
 accelerated gradient methods
 graph, 102, 182
 objective function based on, 103, 182

 heavy-ball method, 55, 57, 65, 68, 71
 Heine-Borel theorem, 209

 image segmentation, 102
 implicit function theorem, 197, 202–203
 incremental gradient method, 77, 95
 cyclic, 80–81
 randomized, 77, 80, 87
 indicator function, 114, 133, 144, 160, 183
 definition of, 144
 proximal operator of, 148
 subdifferential of, 144, 145
 iterate averaging, *see* averaging of iterates

 Jacobian matrix, 188, 196, 198, 202
 Jensen's inequality, 85, 106, 202

 Kaczmarz method
 deterministic, 82–84
 linear convergence of, 83
 randomized, 75, 82–84, 86–87, 91–92, 95
 Karush-Kuhn-Tucker (KKT) conditions, 206
 Kullback-Liebler (KL) divergence, 46
 Kurdyka-Łojasiewicz (KL) condition, 51, 116

 label, 2, 3, 10, 11, 192
 Lagrange multiplier, 172, 182, 184, 196
 Lagrangian, *see* Lagrangian function
 Lagrangian function, 170, 172, 175, 196
 augmented, 180, 181, 183, 184, 186
 for semidefinite program, 173
 Lanczos method, 44
 law of iterated expectation, 88
 learning rate, *see* steplength
 least squares, 4–5, 75, 102, 114
 with zero loss, 82, 91
 level set, 35, 104, 105, 147
 limiting feasible directions, 208–209
 line search, 105
 backtracking, 41–42, 124–125
 exact, 39, 107, 110
 extrapolation-bisection, 40–41, 204–205
 linear independence, 69
 linear programming, 186, 205–206
 simplex method, 206

 Lipschitz constant for gradient, 17, 23, 28, 33,
 38, 76, 87, 88, 101, 104, 122, 123, 125,
 128, 161–163
 componentwise, 104, 113, 115, 165
 componentwise, for quadratic functions,
 104
 for quadratic functions, 104
 Lipschitz constant for Hessian, 43
 Lipschitz continuity, 17
 logistic regression, 9–10, 86
 binary, 9
 multiclass, 10, 12, 192
 loss function, 2, 79, 101
 hinge, 79, 132, 139
 low-dimensional subspace, 2, 3
 lower-semicontinuous function, 134, 144
 Lyapunov function, 55
 for Nesterov's method, 61–68, 71

 matrix optimization, 2, 5–6
 low-rank matrix completion, 5, 114
 nonnegative matrix factorization, 6, 114
 maximum likelihood, 4, 9, 10, 13
 method of multipliers, *see* augmented
 Lagrangian method
 min-max problem, *see* saddle point problem
 minimizer
 global, 15, 29
 isolated local, 15
 local, 15, 148
 strict local, 15, 20
 unique, 15, 147
 minimum principle, 121, 210
 mirror descent, 44–50, 89
 convergence of, 47–50
 missing data, 3
 momentum, 55, 72, 94
 Moreau envelope, 133, 150–151
 gradient of, 150
 relationship to proximal operator, 150

 negative-curvature direction, 43, 44
 nested composition of functions, 188
 Nesterov's method, 55, 57, 70
 convergence on strongly convex functions,
 62–65
 convergence on strongly convex quadratics,
 58–62
 convergence on weakly convex functions,
 66–68

neural networks, 11–13, 132, 188, 191–192
 activation function, 11, 198
 classification, 12
 layer, 11
 parameters, 12
 training of, 12
 Newton's method, 37
 nonlinear equations, 196, 202
 nonnegative orthant, 121, 177, 185
 nonsmooth function, 75, 132–150
 eigenvalues of symmetric matrix, 133
 norms, 133
 normal cone, 48, 133, 144, 175, 208, 212–213
 definition of, 118
 illustration of, 119
 of intersection of closed convex sets,
 144–146
 nuclear norm, 5
 operator splitting, 182
 optimal control, 188, 197–199
 optimality conditions, 133, 209
 for composite nonsmooth function, 146–148
 for convex functions, 134
 examples of, 176–178
 first-order, 196
 first-order necessary, 18–20, 27, 118, 119,
 174–178
 first-order sufficient, 22, 34, 119, 123, 146,
 176, 208
 geometric (for constrained optimization),
 48, 118–120, 123, 146, 174–178
 second-order necessary, 18–20, 42
 second-order sufficient, 20
 order notation, 16, 201
 overfitting, 3
 penalty function, 4
 quadratic, 45, 170–171
 penalty parameter, 171
 perceptron, 78, 80, 95
 as stochastic gradient method, 78
 Polyak–Łojasiewicz (PL) condition, 51, 115,
 213
 prediction, 2
 primal problem, 170, 173, 178
 probability distribution, 75, 79, 202
 progressive function, 190, 191, 195–196
 projected gradient method, 114, 122–127, 130,
 161, 186
 alternative search directions, 126–127
 with backtracking, 124–125

definition of, 122
 short-step, 123–124
 for strongly convex function, 125–126
 projection operator, 120–122, 128, 148, 170,
 185, 210
 nonexpansivity of, 121, 126
 proper convex function, 134
 closed, 134, 148
 prox-operator, *see* proximal operator
 proximal operator, 133, 148–150, 160, 162
 of indicator function, 148
 nonexpansivity of, 149, 161
 of zero function, 149
 proximal point method, 154, 167–168, 180
 and augmented Lagrangian, 180
 definition of, 167
 sublinear convergence of, 167–168
 proximal-gradient method, 110, 126, 148, 149,
 154, 160–164, 168
 linear convergence of, 161–162
 sublinear convergence of, 162
 quadratic programming, 185–186
 OSQP solver, 186
 regression, 2, 79, 101
 regularization, 3
 ℓ_1 , 4, 9
 ℓ_2 , 4, 168
 group-sparse, 10
 regularization function, 3, 13, 26, 101, 103,
 149, 160, 161
 block-separable, 113, 114
 separable, 101, 110, 115, 154, 165
 regularization functions
 block-separable, 116
 regularization parameter, 3, 7, 9, 101, 160
 regularized optimization, *see* composite
 nonsmooth function
 regularizer, *see* regularization function
 restricted isometry property, 6
 robustness, 7
 saddle point problem, 171, 180
 sampling, 79
 with replacement, 113
 without replacement, 113
 semidefinite programming, 173
 separable function, 183, 184
 separating hyperplane, 7, 200, 207, 209, 211,
 212
 separation, 200, 209–212
 of closed convex sets, 210–211

- of hyperplane from convex set, 211–212
- of point from convex set, 209–210
- proper, 211, 212
- strict, 143, 209–211
- set
 - affine, 200
 - affine hull of, 200
 - closure of, 200
 - interior of, 200
 - multiplication by scalar, 200
 - relative interior of, 175, 200, 211
- Sion's minimax theorem, 180
- slack variables, 185
- softmax, 10–12, 14
- solution
 - global, 16, 21, 118, 119
 - local, 16, 21, 118, 119
- spectral radius, 58
- stationary point, 20, 27, 29, 34, 36, 195, 196
- steepest-descent method, 27–33, 43, 44, 55, 62, 68, 76, 77, 101, 105, 111, 149, 153, 155, 160, 161
 - short-step, 28–30, 38, 109, 110
- steplength, 27, 28, 33, 38–42, 78, 110, 122, 161
 - constant step norm, 158
 - decreasing, 93, 158–160
 - exact, 39
 - fixed, 28, 38, 92, 105, 107, 111, 158, 161–163, 167
 - in mirror descent, 49–50
 - for steepest-descent method, 28
 - for subgradient method, 158–160, 180
 - Wolfe conditions and, 39–42
- stochastic gradient descent (SGD), *see* stochastic gradient method
- stochastic gradient method, 38, 75–95, 157, 192, 214
 - accelerated, 96
 - additive noise model, 76, 86
 - basic step, 75–76
 - bounded variance assumption, 85
 - contrast with steepest-descent method, 76
 - convergence analysis of, 87–93
 - epochs, 92–94
 - hyperparameters, 93, 94
 - linear convergence of, 90–92
 - minibatches, 94–95, 192, 199
 - momentum, 94–95
 - parallel implementation, 94
 - SAG, 96
 - SAGA, 96
 - steplength, 81, 85, 88, 90–93
 - sublinear convergence of, 82
 - SVRG, 96
 - variance reduction, 94
- subdifferential, 132–144, 153
 - calculus of, 141–144
 - Clarke, 198
 - closedness and convexity of, 134
 - compactness of, 136, 143
 - definition of, 134
 - and directional derivatives, 138–141
- subgradient, 132–144, 153, 211
 - definition of, 134
 - existence of, 135
 - minimum-norm, 154–156
 - of smooth function, 137
 - and supporting hyperplane of epigraph, 135
- subgradient descent method, 155–156
- subgradient method, 154, 156–160, 179, 198
 - with constant step norm, 158
 - with decreasing steplength, 158–160
 - dual, 179–181, 183, 185
 - with fixed steplength, 158
 - sublinear convergence of, 157–160
- sufficient decrease condition, 39, 41, 125
- support vector machines, 6–9, 78, 79, 132
 - kernel, 9
 - maximum-margin, 7
- supporting hyperplane, 135, 211
- symmetric over-relaxation, 111
- Taylor series, *see* Taylor's theorem
- Taylor's theorem, 15–18, 20, 22–24, 27, 28, 36, 40, 42, 43, 45, 106, 119, 125, 128, 139, 161
 - statement of, 16
 - for vector functions, 202
- telescoping sum, 49, 164
- theorems of the alternative, 205–207
- three-point property, 46, 48
- thresholding
 - hard, 150
 - soft, 150
- topic modeling, 102
- training, 1, 192
- unbiased gradient estimate, 75, 77
- utility maximization, 184–185
- warm start, 168, 171
- Wolfe conditions
 - strong, 53
 - weak, 39–40, 204–205