

## INDEX

Primary references and definitions are indicated in bold.

## A

a posteriori methods, 270, 272, 274  
 a priori methods, 270, 273  
 acquisition function,  $\alpha(x; \mathcal{D})$ , 11, **88**, 94, 96, 98, 150, 247, *see also* expected improvement; knowledge gradient; mutual information; probability of improvement; upper confidence bound  
 batch,  $\beta(x; \mathcal{D})$ , 252  
 gradient, 158, 208  
 optimization, 207  
 action space,  $\mathcal{A}$ , **90**, 245  
 for batch observations, 252  
 for dynamic termination, 104  
 for multifidelity optimization, 263  
 for optimization with fixed budget, 91  
 for sequential procedures, 281  
 for terminal recommendation, 250  
 for terminal recommendations, 110  
 active learning, 136, 275  
 active search, 115, **283**  
 adaptivity gap, 254  
 additive decomposition, **63**, 83  
 aerospace engineering, applications in, 325  
 algorithm configuration, 246, 327  
 anytime algorithm, 209, 239, 248  
 approximate dynamic programming, **100**, 150, 284  
 augmented Chebyshev scalarization, 274  
 automated machine learning (automl), 263, 328, *see also* hyperparameter tuning  
 automatic relevance determination (ARD), 57, 62, 241  
 automobile engineering, applications in, 326

**B**  
 bandits, *see* multi-armed bandits  
 batch observations, 252, 262  
 connection to sequential observations, 254  
 batch rollout, **103**, 152, 154, 284, *see also* rollout  
 Bayes' theorem, 7  
 Bayesian decision theory, 10, **89**, 124  
 isolated decisions, 90

sequential decisions  
 dynamic termination, 103  
 fixed budget, 91  
 multi-armed bandits, 143  
 Bayesian inference  
 introduction to, 6  
 of objective function, 8  
 Bayesian information criterion, 79, 83  
 Bayesian neural networks, 198, 281  
 Bayesian Occam's razor, 71  
 Bayesian quadrature, 33, 278  
 Bayesian regret,  $\mathbb{E}[r_T]$ ,  $\mathbb{E}[R_T]$ , 218, 224, 240  
 Bellman equation, *see* Bellman optimality  
 Bellman optimality, 94, 99  
 beta warping, 58  
 BINOCULARS algorithm, 153  
 biology, applications in, 318  
 biomedical engineering, applications in, 320  
 Bochner's theorem, 50  
 branch and bound, 232, 240

## C

central composite design, 76, 81  
 certainty equivalent, 111  
 characteristic length scale, *see* length scale  
 chemistry, applications in, 288, 313  
 chemoinformatics, 314  
 Cholesky decomposition, 201, 258  
 low-rank updates, 202  
 civil engineering, applications in, 323  
 CMA-ES algorithm, 208  
 combinatorial optimization, 209  
 compactness of domain, 34  
 conditional entropy,  $H[\omega | \mathcal{D}]$ , 115, *see also* entropy  
 conditional mutual information,  $I(\omega; \psi | \mathcal{D})$ , 137, *see also* mutual information  
 confidence interval, 225, 233, 237  
 conformational search, 315  
 conjugate gradients, 203  
 constant liar heuristic, 256  
 constrained optimization, **249**  
 constraint functions, 249  
 unknown, 249  
 constraints on objective function, 36, 39, 56  
 continuity in mean square, 29, *see also* sample path continuity

- continuous differentiability, 31, 221
  - cost-aware optimization, 103, 245, 253, 265
  - covariance function, *see* prior covariance function
  - cross-covariance function, 19, 23, 24, 27, 30, 201, 264
  - cumulative regret,  $R_T$ , 145, 214
  - cumulative reward, 114, 142, 155, 215, 283
  - curse of dimensionality, 61, 208, 285
- D**
- de novo* design, 314
  - decoupled constraint observations, 252
  - deep kernel, 59, 61
  - deep neural networks, ix, 1, 59, 61, 292
  - density ratio estimation, 197
  - design and analysis of computer experiments (DACE), 290
  - determinantal point process, 261, 285
  - differentiability in mean square, 30, *see also* continuous differentiability
  - differential entropy,  $H[\omega]$ , *see* entropy
  - dilation, 56
  - DIRECT algorithm, 208
  - disjoint union, 27
  - drug discovery, *see* molecular design
  - dynamic termination, *see* termination decisions
- E**
- early stopping, 210, 281
  - electrical engineering, applications in, 324
  - elliptical slice sampling, 38
  - embedding, *see* linear embedding, neural embedding
  - entropy search, *see* mutual information
  - entropy,  $H[\omega]$ , 115, *see also* conditional entropy
  - environmental variables, 277
  - expectation propagation, 39, 182, 190, 273, 302
  - expected gain per unit cost, 248
  - expected hypervolume improvement (EHVI), 271
  - expected improvement,  $\alpha_{EI}$ , 81, 95, 113, 117, 127, 151, 158, 193, 196, 199, 265, 266, 268, 286, *see also* simple reward
    - augmented, 166
    - batch, 259
    - comparison with probability of improvement, 132
    - computation with noise, 160
    - alternative formulations, 163
    - gradient, 308
  - computation without noise, 159
    - gradient, 159
  - convergence, 217
    - modified, 154
    - origin, 289
    - worst-case regret with noise, 239, 243
  - expected utility, 90, 93
  - exploration bonus, 145
  - exploration vs. exploitation dilemma, 11, 83, 123, 128, 131, 133, 143, 145, 146, 148, 154, 159, 214, 293
  - exponential covariance function,  $K_{M^{1/2}}$ , 52, 219
  - extreme value theorem, 34
- F**
- factor graph, 37
  - Fano's inequality, 231
  - feasible region,  $\mathcal{F}$ , 249
  - figure of merit, *see* acquisition function
  - fill distance,  $\delta_x$ , 238
  - Fourier transform, 50, 53, 236
  - freeze–thaw Bayesian optimization, 281
  - fully independent training conditional (FITC) approximation, 207
- G**
- Gauss–Hermite quadrature, 172, 258
    - gradient, 309
  - Gaussian process (GP),  $\mathcal{GP}(f; \mu, K)$ , 8, 16, 95, 124, *see also* prior mean function; prior covariance function
    - approximate inference, 35, *see also* sparse spectrum approximation; sparse approximation
    - classification, 41, 283
    - computation of policies with, 157
    - continuity, 28
    - credible intervals, 18
    - differentiability, 30
    - exact inference, 19
      - additive Gaussian noise, 20, 23
      - computation, 201
      - derivative observations, 32
      - exact observation, 22
      - interpretation of posterior moments, 21
    - joint, *see* joint Gaussian process
    - marginal likelihood  $p(\mathbf{y} \mid \mathbf{x}, \theta)$ , 72, 202, 220

- gradient, 307
  - maxima, existence and uniqueness, 33
  - mixture, 38, 75, 193
  - model assessment, selection, and averaging, 67
  - modeling, 45
  - posterior predictive distribution, 25, 157, 208
    - gradient, 307
    - sampling, 18
  - gene design, 319
  - generalized additive models, 63
  - generalized linear models, 41
  - GLASSES algorithm, 152
  - global reward, 113, 117, 129, 172, *see also* knowledge gradient
  - GP-SELECT algorithm, 285
  - Gram matrix,  $K(\mathbf{x}, \mathbf{x})$ , 17, 49
  - grid search, 3, 240
- H**
- Hamiltonian Monte Carlo (HMC), 75
- Heine–Borel theorem, 30
- heteroskedastic noise, 4, 23, 166
- Hölder continuity, 35, 221, 240
- horizon, 93, 125, 151, 245, 254
- hubris of youth, ix
- human–computer interfaces, 329
- hyperband algorithm, 282
- hyperparameter tuning, 1, 61, 109, 263, 267, 291, 327
- hyperparameters,  $\theta$ , 68, *see also* length scale; output scale
  - unknown, effect of convergence, 241
- I**
- ill-conditioning, 203
- incumbent value,  $\phi^*$ , 128, 159, 251
- inducing values,  $\mathbf{v}$ , 205
- infill function, *see* acquisition function
- information capacity,  $\gamma_{\tau}$ , 221, 225, 229, 233, 237
  - bounds, 222
- information gain, 115, 117, 135, 180, 187, *see also* mutual information
- initialization, 210
- instantaneous regret,  $\rho_{\tau}$ , 214, 241
- integrated expected conditional improvement, 251
- intrinsic dimensionality, 61
- inverse problems, 318, 319
- isotropy, 50, 54
- iterative numerical methods, 203
- Iverson bracket, xiii
- J**
- joint Gaussian process, 27, 267, *see also* cross-covariance function
  - between function and gradient, 30
  - exact inference, 28
  - for multifidelity modeling, 264
  - marginals, 27
- K**
- kernel, *see* prior covariance function
- kernel kernel, 83
- knowledge gradient,  $\alpha_{\text{KG}}$ , 113, 129, 172, 193, 266, 276, *see also* global reward
  - batch, 259
  - computation, 172
  - discrete domain, 173
  - gradient, 310
  - KGCP approximation, 175
  - origin, 290
- Kolmogorov extension theorem, 16
- kriging believer heuristic, 256
- Kullback–Leibler divergence,  $D_{\text{KL}}[p \parallel q]$ , 115, 137, 206
- L**
- Laplace approximation, 39, 41, 76, 79, 83, 301
- learning curve, 281
- length scale, 54, 56, 69, *see also* automatic relevance determination
- likelihood, 7, *see also* observation model
  - marginal, *see* marginal likelihood
- limited lookahead, *see* lookahead
- line search methods, 285
- linear covariance function,  $K_{\text{LIN}}$ , 54
- linear embedding, 58, 62, 209
- linear scalarization, 274
- linear transformations
  - of domain, 55, 56, 62
  - of Gaussian processes, 33, *see also* Bayesian quadrature; joint Gaussian process: between function and gradient
- Lipschitz continuity, 227, 240, 258
- local optimization, 285
- local optimum, conditioning a Gaussian process on a, 36, 182
- local penalization, 257
- lookahead, 101, 125, 150, 284, *see also* one-step lookahead; two-step lookahead

- low-dimensional structure, 58, 61, 62, 209, 294
- low-discrepancy sequence, 177, 211, 239
- lower regret bounds
- Bayesian regret with noise, 230
  - Bayesian regret without noise, 240
  - worst-case regret with noise, 236
  - worst-case regret without noise, 239
- Löwner order, 22, 48
- M**
- Mahalanobis distance, 48, 57, 297
- manifold Gaussian process, 59, 61
- marginal likelihood,  $p(y \mid \mathbf{x}, \theta)$ , 71
- materials science, applications in, 263, 313
- Matérn covariance function,  $K_M$ , 51, 124, 221, 238, 241
- matrix calculus convention, xiii
- maximum a posteriori inference, *see* model selection
- maximum information gain, *see* information capacity
- max-value entropy search,  $\alpha_{MES}$ , 187
- for multiobjective optimization, 273
  - gradient, 191
- mean function, *see* prior mean function
- mechanical engineering, applications in, 325
- model assessment, 45, 67, 70, *see also* model posterior
- model averaging, 74, 116
- in acquisition function, 192
  - multiple model structures, 79
- model evidence, *see* marginal likelihood
- model posterior,  $p(\theta \mid \mathcal{D})$ , 71
- model prior,  $p(\theta)$ , 70
- model selection, 73, 329
- multiple model structures, 79
- model space, *see* model structure
- model structure,  $\mathcal{M}$ , 68, *see also* hyperparameters
- posterior,  $\Pr(\mathcal{M} \mid \mathcal{D})$ , 79
  - prior,  $\Pr(\mathcal{M})$ , 78
  - search, 81
- model,  $p(y \mid \mathbf{x})$ , 68
- molecular design, 62, 209, 283, 313
- molecular fingerprint, 314
- Monte Carlo sampling, 37, 75, 84, 181, 187, 258, 273
- multi-armed bandits, 141
- infinite-armed, 144
  - optimal policy, 143
  - origin, 292
- multifidelity optimization, 26, 263, 321
- multiobjective optimization, 26, 269
- multitask optimization, 266
- mutual information,  $I(\omega; \psi)$ , 116, 135, 291, *see also* information gain; conditional mutual information
- with  $f^*$ ,  $\alpha_{f^*}$ , 140, 187, 193, 266, 291, *see also* output-space predictive entropy search; max-value entropy search
  - with  $x^*$ ,  $\alpha_{x^*}$ , 139, 180, 193, 266, 291, *see also* predictive entropy search
- myopic approximation, *see* lookahead
- N**
- nats, xiii
- needle in a haystack analogy, 216, 236, 239, 240
- neural architecture search, 328
- neural embedding, 59, 61, 209
- neural networks, 198
- no U-turn sampler (NUTS), 76
- nonmyopic policies, 150, 284, 294
- no-regret property, 145, 215
- Nyström method, 207
- O**
- objective function posterior,  $p(f \mid \mathcal{D})$ , 9, 74, 92, *see also* Gaussian process: exact inference, approximate inference
- model-marginal, *see* model averaging
- objective function prior,  $p(f)$ , 8, *see also* Gaussian process
- observation costs, 104
- unknown, 245
- observation model,  $p(y \mid x, \phi)$ , 4
- additive Gaussian noise, 4, 23, 69, 78, 157
  - additive Student- $t$  noise, 36, 38, 41, 282
  - exact observation, 4, 22
  - for unknown costs, 246
- observation noise scale,  $\sigma_n$ , 4, 23, 69, 78, 157, 203, 222
- one-step lookahead, 94, 102, 126, 171, 245, 247, 251, 252, 283, 289
- cost-aware optimization, 106
  - with cumulative reward, 155
  - with global reward, *see* knowledge gradient
  - with information gain, *see* mutual information

- with simple reward, *see* expected improvement
  - optimal design, 287
  - optimal policy
    - batch observations, 253
    - computational cost, 99, 125, 284
    - generic, 245
    - multi-armed bandits, 143
    - sequential optimization, 98
  - optimism in the face of uncertainty, 145
  - optimization policy, 3, *see also* acquisition function; grid search; optimal policy; random search; Thompson sampling
    - optimal, *see* optimal policy
  - Ornstein–Uhlenbeck (ou) process, 52, 174, 290
  - output scale, 55, 69, 242
  - output-space predictive entropy search,
    - $\alpha_{\text{OPES}}$ , 187
    - gradient, 311
- P**
- PareGO algorithm, 275
  - Pareto dominance, 270
  - Pareto frontier, 153, 269
  - Pareto optimality, *see* Pareto frontier
  - Parzen estimation, 197
  - periodic covariance function, 35, 58, 60
  - physics, applications in, 317
  - plant breeding, 1, 320
  - posterior distribution, 7
  - posterior predictive distribution,  $p(y \mid x, \mathcal{D})$ ,
    - 8, 25, 74, 92, 157, 208, 283
    - for multifidelity modeling, 264
    - for unknown costs, 246
    - model-marginal, *see* model averaging
  - predictive entropy search,  $\alpha_{\text{PES}}$ , 180
    - batch, 260
    - for multiobjective optimization, 273
    - gradient, 311
  - preference optimization, 282, 330
  - prior covariance function,  $K(x, x')$ , 17, 49, 67,
    - see also* exponential, linear, Matérn, spectral mixture, squared exponential covariance functions; automatic relevance determination
    - addition, 60, 63
    - multiplication, 60
    - scaling, 55
    - warping, 56
  - prior distribution, 6
  - prior mean function,  $\mu(x)$ , 16, 46, 67
    - concave quadratic, 48
    - constant, 47, 69, 124, 207
    - marginalization of parameter, 47, 69
    - impact on posterior mean, 46
    - linear combination of bases, 48
  - probability of improvement,  $\alpha_{\text{PI}}$ , 131, 167, 193,
    - 196, 199
    - batch, 260
    - comparison with expected improvement, 132
    - computation with noise, 169
    - gradient, 309
    - computation without noise, 167
    - gradient, 169
    - convergence, 217
    - correspondence with upper confidence bound, 170
    - origin, 289
    - selection of improvement target, 133, 289
  - protein design, 319
  - pseudopoints, *see* inducing values
- Q**
- quantile function,  $q(\pi)$ , 145, 165, 170
- R**
- random embedding, 63
  - random forests, 196
  - random search, 3
  - reaction optimization, 288, 315
  - regret, 213, *see also* simple regret; cumulative regret; Bayesian regret; worst-case regret
  - reinforcement learning, applications in, 322
  - representer points, 180, 188
  - reproducing kernel Hilbert space (RKHS),
    - $\mathcal{H}_K$ , 219, 224, 232, 242
    - RKHS ball,  $\mathcal{H}_K[B]$ , 220, 224, 232
    - RKHS norm,  $\|f\|_{\mathcal{H}_K}$ , 220
  - risk neutrality, 111, 127, 250
  - risk tolerance, 111
  - risk vs. reward tradeoff, 112, 269
  - robotics, applications in, 277, 282, 321
  - robust optimization, 277, 322
  - rolling horizon, 101
  - rollout, 102, 151, 263
    - batch, *see* batch rollout
- S**
- $\mathcal{S}$  metric, 271

- safe optimization, 322
  - sample path continuity, 30, 34, 218, 221
  - scalarization, 273
  - second-derivative test, 36, 40
  - separable covariance function, 264
  - sequential analysis, 287
  - sequential experimental design, 288
  - sequential simulation, 255
  - signal variance, *see* output scale
  - simple regret,  $r_\tau$ , 214, 215, 216, 230, 237
  - simple reward, 95, 112, 117, 127, 158, 165, 251, 284, *see also* expected improvement
  - small data, 68
  - sparse approximation, 204
  - sparse spectrum approximation, 51, 178
  - spectral density,  $\kappa$ , 51, 53
  - spectral measure,  $\nu$ , 51, 53
  - spectral mixture covariance function,  $K_{SM}$ , 51, 53
  - spectral points, 178
  - sphere packing, 238
  - squared exponential covariance function,  $K_{SE}$ , 17, 52, 221
  - stationarity, 50, 56, 58, 178, 207, 236
  - stochastic gradient ascent, 286
  - stochastic optimization, 277
  - stochastic process, 8, 16, *see also* Gaussian process
  - structural search, 316
  - Student- $t$  process, 282
  - sub-Gaussian distribution, 233
- T**
- terminal recommendations, 90, 109, 118
  - termination decisions, 5, 253
    - optimal, 103
    - practical, 211
  - termination option,  $\emptyset$ , 104
  - Thompson sampling, 148, 176, 181, 187, 195, 259
    - acquisition function view, 148
    - batch, 261
    - computation, 176
    - origin, 292
    - regret bounds
      - Bayesian regret with noise, 229
      - worst-case regret with noise, 233
- U**
- upper confidence bound,  $\alpha_{UCB}$ , 145, 170, 195, 266
    - batch, 261
    - computation, 170
    - correspondence with probability of improvement, 170
    - gradient, 170
    - origin, 289
    - regret bounds
      - Bayesian regret with noise, 225
      - worst-case regret with noise, 233, 243
    - selecting confidence parameter, 147
  - utility function
    - for active search, 283
    - for constrained optimization, 250
    - for cost-aware optimization, 104
    - for isolated decisions,  $u(a, \psi, \mathcal{D})$ , 90
    - for multifidelity optimization, 264
    - for multitask optimization, 268
    - for optimization,  $u(\mathcal{D})$ , 93, 109, 245, *see also* cumulative reward; global reward; simple reward; information gain
    - for terminal recommendations,  $v(\phi)$ , 111
- V**
- value of data,  $a_\tau^*$ , 95, 101
  - value of sample information, 126
  - variational inference, 39, 206
  - virtual screening, 314
  - von Neumann–Morgenstern theorem, 90, 120
- W**
- weighted Euclidean distance, 57, *see also* automatic relevance determination
  - Wiener process, 174, 217, 232, 242, 290
  - wiggleness, 56, 198, *see also* length scale
  - worst-case regret,  $\bar{r}_\tau, \bar{R}_\tau$ , 218, 232, 239