Stochastic and deterministic algorithms for continuous black-box optimization
Wang, H.

Citation

Version: Not Applicable (or Unknown)
License: Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from: https://hdl.handle.net/1887/66671

Note: To cite this publication please use the final published version (if applicable).
The handle [http://hdl.handle.net/1887/66671](http://hdl.handle.net/1887/66671) holds various files of this Leiden University dissertation.

**Author:** Wang, H.  
**Title:** Stochastic and deterministic algorithms for continuous black-box optimization  
**Issue Date:** 2018-11-01
Gaussian Distribution

Assume the probability space \((\Omega, \mathcal{F}, \mathbb{P})\) and a measurable space \((\mathbb{R}, \mathcal{B})\), where \(\mathcal{B}\) is the Borel algebra on \(\mathbb{R}\). A random variable \(X : \Omega \to \mathbb{R}\) is said to be normally distributed if and only if its probability distribution \(\mathbb{P}_X : \mathcal{B} \to [0, 1]\), defined as a push-forward measure, \(\forall B \in \mathcal{B}, \mathbb{P}_X(B) := \mathbb{P}(X^{-1}(B))\), admits the following form:

\[
\mathbb{P}_X(B) = \int_B \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x - m)^2}{2\sigma^2} \right) \, d\lambda,
\]

where \(\lambda\) is the Lebesgue measure on \(\mathbb{R}\) and \(m, \sigma^2\) are the mean and variance of \(X\), respectively. We typically use the notation \(X \sim \mathcal{N}(m, \sigma^2)\). This distribution \(\mathbb{P}_X\) is called Gaussian measure and the notation \(G_{m, \sigma^2}\) is assigned to it. The cumulative distribution function (c.d.f.) of \(X\) is

\[
\Phi_{m, \sigma^2}(x) = G_{m, \sigma^2}(\{X \in \mathbb{R} : X \leq x\}).
\]

In addition, the probability density function (p.d.f.) of \(X\) is the Radon-Nikodym derivative of \(G_{m, \sigma^2}\) w.r.t. \(\lambda\):

\[
\phi_{m, \sigma^2}(x) = \frac{dG_{m, \sigma^2}}{d\lambda} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x - m)^2}{2\sigma^2} \right),
\]

that is, by definition, the integrand in Eq. (A.1). In the multivariate case, consider the measurable space \((\mathbb{R}^n, \mathcal{B}^n)\) where \(\mathcal{B}^n\) is the Borel algebra on \(\mathbb{R}^n\). A random vector \(\mathbf{x} = (X_1, X_2, \ldots, X_n)\top : \Omega \to \mathbb{R}^n\) is said to follow the multivariate Gaussian distribution, if and only if any linear combination \(c^\top \mathbf{x}, c \in \mathbb{R}^n\) admits the distribution as in Eq. (A.1). In addition, the distribution of \(\mathbf{x}\) is

\[
\forall B \in \mathcal{B}^n, G_{m, \mathbf{K}}(B) = \int_B (2\pi)^{-\frac{n}{2}} \det(\mathbf{K})^{\frac{1}{2}} \exp \left( -\frac{1}{2} (\mathbf{y} - \mathbf{m})^\top \mathbf{K}^{-1} (\mathbf{y} - \mathbf{m}) \right) \, d\lambda^n,
\]

that is, by definition, the integrand in Eq. (A.1). In the multivariate case, consider the measurable space \((\mathbb{R}^n, \mathcal{B}^n)\) where \(\mathcal{B}^n\) is the Borel algebra on \(\mathbb{R}^n\). A random vector \(\mathbf{x} = (X_1, X_2, \ldots, X_n)\top : \Omega \to \mathbb{R}^n\) is said to follow the multivariate Gaussian distribution, if and only if any linear combination \(c^\top \mathbf{x}, c \in \mathbb{R}^n\) admits the distribution as in Eq. (A.1). In addition, the distribution of \(\mathbf{x}\) is

\[
\forall B \in \mathcal{B}^n, G_{m, \mathbf{K}}(B) = \int_B (2\pi)^{-\frac{n}{2}} \det(\mathbf{K})^{\frac{1}{2}} \exp \left( -\frac{1}{2} (\mathbf{y} - \mathbf{m})^\top \mathbf{K}^{-1} (\mathbf{y} - \mathbf{m}) \right) \, d\lambda^n,
\]
where \( \lambda^n \) is the \( n \)-dimensional Lebesgue measure on \((\mathbb{R}^n, \mathcal{B}^n)\) and \( \mathbf{m}, \mathbf{K} \) are the mean and covariance matrix of \( \mathbf{x} \). As with the univariate case, we shall take the notation \( \mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{K}) \) and its cumulative distribution function is,

\[
\Phi_{\mathbf{m}, \mathbf{K}}(\zeta) = \mathcal{G}_{\mathbf{m}, \mathbf{K}}(\{ \mathbf{x} \in \mathbb{R}^n : \mathbf{x} \leq \zeta \}).
\]

Given an arbitrary partition on \( \mathbf{x} = (x_1^T, x_2^T)^T \), in which \( x_1, x_2 \) have \( n_1 \) and \( n_2 \) components, respectively. The distribution of \( \mathbf{x} \) can be re-written as

\[
\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \right),
\]

where all sub-mean vectors and sub-covariance matrices are obtained by applying the same partition on \( \mathbf{m} \) and \( \mathbf{K} \). The marginal distribution of \( x_1 \) is Gaussian:

\[
x_1 \sim \mathcal{N}(\mathbf{m}_1, \mathbf{K}_{11}). \tag{A.3}
\]

The result holds for \( x_2 \) in the same manner. In addition, the conditional distribution of \( x_1 \) on \( x_2 = \mathbf{v} \) is Gaussian [Tong 2012]:

\[
x_1 \mid x_2 = \mathbf{v} \sim \mathcal{N}(\mathbf{m}_1 + \mathbf{K}_{12} \mathbf{K}_{22}^{-1}(\mathbf{v} - \mathbf{m}_2), \mathbf{K}_{11} - \mathbf{K}_{12} \mathbf{K}_{22}^{-1} \mathbf{K}_{21}) . \tag{A.4}
\]

Often, the value of Gaussian random variables is restricted:

\[
X \sim \mathcal{N}(m, \sigma^2), \quad X_R = \max\{0, X\}.
\]

The random variable \( X_R \) is known as the Rectified Gaussian and its distribution shall be denoted as \( \mathcal{N}_R(m, \sigma^2) \). Note that the rectification “concentrates” all the probability measure in \((-\infty, 0)\) to the rectification point 0, leading to an infinite impulse at this point. Thus, the p.d.f. of \( X_R \) is:

\[
p_{X_R}(x) = \Phi_{m, \sigma^2}(0) \delta(x) + \phi_{m, \sigma^2}(x) H(x), \tag{A.5}
\]

where \( \delta \) is the Dirac delta (distribution)\(^1\) and \( H \) is the step function:

\[
\delta(x) = \begin{cases} 
\infty & x = 0, \\
0 & x \neq 0
\end{cases}, \quad H(x) = \begin{cases} 
0 & x \leq 0, \\
1 & x > 0
\end{cases}.
\]

The rectification is sometimes confused with the so-called truncated Gaussian, which is the distribution of a Gaussian variable \( X \sim \mathcal{N}(m, \sigma^2) \) conditioning on an interval \((a, b) \subset \mathbb{R} \):

\[
p(x \mid a < X < b) = \frac{\phi_{m, \sigma^2}(x)}{\Phi_{m, \sigma^2}(b) - \Phi_{m, \sigma^2}(a)}.
\]

\(^1\)Formally, the Dirac delta should be defined either as a distribution or measure. We use the heuristic characterization here for the sake of simplicity.
Proof

B.1 Theorem 5.3

Proof. Let us define \( a := -\nabla f_1^{(2)} \) and \( b := \nabla f_2^{(1)} \), such that \( \tilde{A}_1 = ba^\top \) and

\[
\nabla^2 H_F(X) = \begin{pmatrix}
D_1 & ba^\top \\
abla_2 & D_2
\end{pmatrix}.
\]

For two block matrices, their column vectors are denoted as: \( D_1 = (d_1, \ldots, d_n) \) and \( D_2 = (d'_1, \ldots, d'_n) \). The hypervolume Hessian is of size \( 2n \times 2n \) and its determinant can be simplified using the Laplace expansion along the first \( n \) rows of the \( \nabla^2 H_F(X) \). To achieve this, \( n \) distinct columns need to be selected out of \( 2n \) rows. Let \( S \) be the set of the \( n \)-element subsets of \( \{1, 2, \ldots, 2n\} \):

\[
S = \{\{1, 2, \ldots, n\}, \{1, 2, \ldots, n-1, n+1\}, \ldots\}
\]

For every \( L \in S \), we define its complement \( L' := \{1, 2, \ldots, 2n\} \setminus L \). Note that a permutation is defined on \( \{1, 2, \ldots, 2n\} \), by appending \( L' \) to \( L \): \( \{L, L'\} \) and we shall use \( N(L) \) to denote the number of inversions in \( \{L, L'\} \). According to the Laplace expansion, such a determinant can be expressed as:

\[
\det (\nabla^2 H_F(X)) = \sum_{L \in S} (-1)^{N(L)} b_L c_{L'},
\]

where \( b_L \) is the cofactor of the hypervolume Hessian, which is the determinant of the minor matrix obtained by keeping the first \( n \) rows and \( n \) columns given in \( L \). Similarly, \( c_{L'} \) is the complementary cofactor of \( b_L \), obtained by removing the first \( n \) rows and \( n \) columns given in \( L \). For example, if \( L = \{1, 2, \ldots, n\} \), then \( b_L = \det(D_1) \) and \( c_{L'} = \det(D_2) \). In particular, when \( L \) contains two or more elements from \( \{2n+1, 2n+2, \ldots, 2n\} \), meaning that at least two columns from
B. PROOF

\( \mathbf{b \mathbf{a}^T} \) are chosen to compute \( b_L \), it is obvious that the cofactor \( b_L \) is zeros because all the columns from \( \mathbf{b \mathbf{a}^T} \) are linear dependent. Using this observation, the expansion can be simplified:

\[
\begin{align*}
\det (\nabla^2 \mathcal{H}_F(\mathbf{X})) &= \det (\mathbf{D}_1) \det (\mathbf{D}_2) _{L=\{1,2,\ldots,n\}} \\
&+ (-1)^1 \det ((\mathbf{d}_1, \ldots, \mathbf{d}_{n-1}, a_1 \mathbf{b})) \det ((b_n \mathbf{a}, \mathbf{d}_2^\prime, \ldots, \mathbf{d}_n^\prime)) _{L=\{1,2,\ldots,n-1,n+1\}} \\
&+ (-1)^2 \det ((\mathbf{d}_1, \ldots, \mathbf{d}_{n-1}, a_2 \mathbf{b})) \det ((b_n \mathbf{a}, \mathbf{d}_1^\prime, \mathbf{d}_3^\prime, \ldots, \mathbf{d}_n^\prime)) + \cdots _{L=\{1,2,\ldots,n-1,n+2\}} \\
&+ (-1)^n \det ((\mathbf{d}_1, \ldots, \mathbf{d}_{n-1}, a_2 \mathbf{b})) \det ((b_n \mathbf{a}, \mathbf{d}_1^\prime, \mathbf{d}_2^\prime, \ldots, \mathbf{d}_{n-1}^\prime)) + \cdots _{L=\{1,2,\ldots,n-1,2n\}}
\end{align*}
\]

There are \( n \) terms shown in the equation above, resulting from choosing the first \( n-1 \) columns and one column from \( \{2n+1, 2n+2, \ldots, 2n\} \). Those terms can also be simplified:

\[
\begin{align*}
&(-1)^1 a_1 b_n \det ((\mathbf{d}_1, \ldots, \mathbf{d}_{n-1}, \mathbf{b})) \det ((\mathbf{a}, \mathbf{d}_2^\prime, \ldots, \mathbf{d}_n^\prime)) \\
&+ (-1)^2 a_2 b_n \det ((\mathbf{d}_1, \ldots, \mathbf{d}_{n-1}, \mathbf{b})) \det ((\mathbf{d}_1^\prime, \mathbf{a}, \mathbf{d}_3^\prime, \ldots, \mathbf{d}_n^\prime)) + \cdots \\
&+ (-1)^{2i-1} a_i b_n \det ((\mathbf{d}_1, \ldots, \mathbf{d}_{n-1}, \mathbf{b})) \det ((\mathbf{d}_1^\prime, \ldots, \mathbf{d}_{i-1}^\prime, \mathbf{a}, \mathbf{d}_{i+1}^\prime, \ldots, \mathbf{d}_n^\prime)) \\
&= -b_n \det ((\mathbf{d}_1, \ldots, \mathbf{d}_{n-1}, \mathbf{b})) \det (\mathbf{D}_2) \sum_{i=1}^{n} a_i \frac{\det ((\mathbf{d}_1^\prime, \ldots, \mathbf{d}_{i-1}^\prime, \mathbf{a}, \mathbf{d}_{i+1}^\prime, \ldots, \mathbf{d}_n^\prime))}{\det ((\mathbf{d}_1^\prime, \mathbf{d}_2^\prime, \ldots, \mathbf{d}_n^\prime))} \\
&= -b_n \det ((\mathbf{d}_1, \ldots, \mathbf{d}_{n-1}, \mathbf{b})) \det (\mathbf{D}_2) \mathbf{a}^\top \mathbf{D}_2^{-1} \mathbf{a}
\end{align*}
\]

Note that the last step above is according to Cramer’s rule for the equation \( \mathbf{D}_2 \mathbf{x} = \mathbf{a} \) (\( \mathbf{D}_1 \) and \( \mathbf{D}_2 \) are assumed to be nonsingular):

\[
x_i = \frac{\det ((\mathbf{d}_1^\prime, \ldots, \mathbf{d}_{i-1}^\prime, \mathbf{a}, \mathbf{d}_{i+1}^\prime, \ldots, \mathbf{d}_n^\prime))}{\det ((\mathbf{d}_1^\prime, \mathbf{d}_2^\prime, \ldots, \mathbf{d}_n^\prime))}
\]
In principle, the same simplification here can be applied to other terms in the hypervolume Hessian determinant:

\[
\text{det}\left(\nabla^2 \mathcal{H}_F(X)\right) = \text{det}(D_1) \text{det}(D_2) \\
- b_n \text{det}\left(\left(\mathbf{d}_1, \ldots, \mathbf{d}_{n-1}, \mathbf{b}\right)\right) \text{det}(D_2) (\mathbf{a}^\top D_2^{-1} \mathbf{a}) \\
- b_{n-1} \text{det}\left(\left(\mathbf{d}_1, \ldots, \mathbf{d}_{n-2}, \mathbf{b}, \mathbf{d}_n\right)\right) \text{det}(D_2) (\mathbf{a}^\top D_2^{-1} \mathbf{a}) - \cdots \\
- b_1 \text{det}\left(\left(\mathbf{b}, \mathbf{d}_2, \ldots, \mathbf{d}_n\right)\right) \text{det}(D_2) (\mathbf{a}^\top D_2^{-1} \mathbf{a}) \\
= \text{det}(D_1) \text{det}(D_2) \left[1 - (\mathbf{a}^\top D_2^{-1} \mathbf{a}) \sum_{i=1}^{n} b_i \frac{\text{det}\left(\left(\mathbf{d}_1, \ldots, \mathbf{d}_{i-1}, \mathbf{b}, \mathbf{d}_{i+1}, \ldots, \mathbf{d}_n\right)\right)}{\text{det}\left(\left(\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n\right)\right)}\right] \\
= \left(1 - (\mathbf{a}^\top D_2^{-1} \mathbf{a}) (\mathbf{b}^\top D_1^{-1} \mathbf{b})\right) \text{det}(D_1) \text{det}(D_2)
\]

Again, in the last step above the same argument as in Eq. (B.1) is applied. Because matrices \(D_1\) and \(D_2\) are nonsingular, the hypervolume Hessian matrix is nonsingular as long as \(1 - (\mathbf{a}^\top D_2^{-1} \mathbf{a}) (\mathbf{b}^\top D_1^{-1} \mathbf{b})\) is not zero."
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


Acronyms

**BLP**  Best Linear Estimator. 50

**BLUE**  Best Linear Unbiased Estimator. 48

**BLUP**  Best Linear Unbiased Predictor. 48

**ECDF**  Empirical Cumulative Distribution Function. 36, 101

**EGO**  Efficient Global Optimization. 37, 90, 113, 114

**EI**  Expected Improvement. 78, 90, 93, 96, 103, 108, 114, 154

**GEI**  Generalized Expected Improvement. 91, 92, 98, 100

**GLS**  Generalized Least Squares. 48

**GPR**  Gaussian Process Regression. 39, 43, 44, 56, 59, 63, 69, 87, 88, 95, 100, 152

**KKT**  Karush-Kuhn-Tucker conditions. 57, 95, 125

**LHS**  Latin Hypercube Sampling. 57, 101

**LUP**  Linear Unbiased Predictor. 47

**MAP**  Maximum a Posterior. 59

**MGF**  Moment-Generating Function. 96

**MGFI**  Moment-Generating Function of Improvement. 98, 100, 101, 103, 107, 154, 155

**MSE**  Mean Squared Error. 37, 39, 49, 50, 69, 71, 73, 80, 82, 94, 108
Acronyms

**RBF** Radial Basis Functions. 46

**RKHS** Reproducing Kernel Hilbert Space. 54, 56, 57, 155
Index

Ackley function, 114
acquisition function, 37, 87
almost everywhere, 120
attainable, 95

basis functions, 45
Bayesian Committee Machines, 65
Bayesian optimization, 37
Bayesian statistics, 58
best linear predictor, 50
best linear unbiased predictor, 48
bi-objective, 92
Branin function, 113

characteristic function, 51
Coefficient of determination, 77
condition number, 53
conditional distribution, 59, 158
Constant Liar, 106
convergence in probability, 4
convergence rate analysis, 27
covariance function, 44
    auto-covariance, 48
covariance matrix, 158
cumulative distribution function, 157, 158
cumulative regret, 88
data generation process, 51
decision space, 3

Efficient Global Optimization, 37, 90
Euclidean ball, 3
Expected Improvement, 37, 90
    Bootstrapped, 90
    Generalized, 91
    Multi-point, 106
    Multiple Generalized, 92
    Weighted, 91

Fuzzy C-means, 67, 74

Gaussian
    Markov Random Fields, 64
    measure, 157
    mixture models, 66
    multivariate, 15, 157
    Process, 43, 58
Gaussian Mixture Model Cluster Kriging, 74, 78
Generalized Least Squares, 48
Global minimum, 3
GPR variance, 59

Hartman6, 113
Hilbert space, 53
Himmelblau’s function, 114
hyper-parameters, 46
hypervolume indicator, 121
    gradient, 119, 128
    Gradient Ascent, 95, 134
INDEX

Hessian matrix, 120
Newton method, 146
improvement, 89
Improvement-based infill criteria, 89
indicator function, 51
Infill Criteria
  parallelization, 105
  parameterized, 88
infill criterion, 37, 87
Isotropy, 46
K-means, 66, 73
Karush-Kuhn-Tucker, 57, 143
Kriging, 43
  Cluster Kriging, 43
  local Kriging, 69
  Ordinary Kriging, 39, 45
  Simple Kriging, 45, 50
  Universal Kriging, 45, 51
Kriging Believer, 106
Kriging MSE, 49, 87
Kriging nugget, 52
Kriging predictor, 53
Kriging RMSE, 49
Lagrange Multiplier, 71
Lagrange Multipliers, 48
Lebesgue measure, 158
likelihood, 58
linear unbiased predictor, 47
Local minimum, 3
local search, 3
Lower Confidence Bound, 37, 88
marginal distribution, 158
Matrix Calculus, 8
Maximum a Posterior Probability, 59
Maximum Likelihood Estimation, 63
mean squared error, 37, 47
Mean Standardized Log Loss, 77
metaheuristics, 5
metric space, 3
Model Tree Cluster Kriging, 74, 78
Moment-Generating Function, 96
of Improvement, 98
multi-objective optimization, 119
Mutation by Optimization, 37
niching evolution strategy, 108
Niching-q-EI, 108
non-informative, 61
nonparametric regression, 51
nugget effect, 52
nugget variance, 52
Optimally Weighted
  Cluster Kriging, 73, 78
  Fuzzy Cluster Kriging, 74
Pareto efficient, 7
  efficient set, 7
Pareto front, 7
Pareto order, 7
positive semi-definite matrix, 47
positive-definite, 50
  positive-definite kernel, 44, 60
  positive-definite matrix, 47
posterior, 58
  kernel, 60
  mean, 60
prior, 58
  mean, 59
probability density function, 157
probability distribution, 157
Probability of Improvement, 37, 90
radial basis functions, 46
INDEX

Radon-Nikodym derivative, 157
random forests, 87
Rastrigin function, 114
Rectified Gaussian, 90, 158
regression function, 51
representer theorem, 56
Reproducing Kernel Hilbert Space, 54
risk function, 47
sample path, 44
sampling error, 15
search space, 3
Semivariogram, 52
separable space, 53
set-oriented numerics, 120
simple random sampling, 15
Sparse On-Line Gaussian Processes, 64
Standardized Mean Squared Error, 77
Stationary, 46
statistical model, 69
Stochastic Optimization, 4
Subset of Data, 64
Subset of Regressors, 64
supermartingale, 4
support vector regression, 87
supremum norm, 55
uniform random orthogonal vectors, 21
Upper Confidence Bound, 88
weakly isotropic, 46
weakly stationary, 46