Gaussian Markov Random Fields for Discrete Optimization via Simulation: Framework and Algorithms

Presented By LI Zaile

zaileli21@m.fudan.edu.cn
Department of Management Science
Fudan University

Peter L. Salemi, Eunhye Song, Barry L. Nelson, Jeremy Staum (2019).
Operations Research 67(1):250-266.

January 5, 2022

Outline

The Problem

$$\text{minimize } y(\mathbf{x}) \triangleq \mathrm{E}[Y(\mathbf{x})] \text{ subject to } \mathbf{x} \in \mathscr{X}$$

The feasible set $\mathscr{X}$ is a finite subset of the d-dimensional integer lattice $\mathbb{Z}^d$ and $n \triangleq |\mathscr{X}|$.

Let $\mathbf{x}^\star$ denote the unknown best of the n feasible solutions, which we assume for simplicity of exposition is unique: $\mathbf{x}^\star = \arg\min_{\mathbf{x}} y(\mathbf{x})$. Let $\tilde{\mathbf{x}}$ be the selected solution by whatever method.

global convergence.

Adaptive random search (ARS)

Commonly, adaptive random search involves

- Surrogate Modeling
- Anticipated spatial structure (Correlation)
- Algorithm Design that can
  - learn spatial relationships by sampling
  - balance exploration and exploitation
  - facilitate a stopping criterion that considers the uncertainty at feasible solutions that have and have not been simulated

For example, the gaussian random field (GRF) => numerical issues with inverse of dense covariate matrix (except the 2014 OR Paper of Jeff).

Why A Discrete Gaussian Markov Random Field (GMRF) can be Great?

- If we are interested in predicting the value of the objective function at a feasible solution, then the values of the objective function at the feasible solutions in a neighborhood of it would often be sufficient; other feasible solutions would provide little additional information.
- GMRFs can be defined on a lattice, so the use of GMRFs in DOvS problems is more natural than using a GRF with a continuous domain.
- The dependence in a GMRF is defined by its precision matrix, which is the inverse of the covariance matrix. Using the Markov structure of GMRFs, the precision matrix can be constructed to be sparse.

Other Works

- P. Salemi, J. Staum and B. L. Nelson, "Gaussian Markov Random Fields for Simulation Metamodeling," Proceedings of the 2013 Winter Simulation Conference, 543-554.
- Salemi P, Nelson BL, Staum J (2014) Discrete optimization via sim- ulation using Gaussian Markov random fields. Tolk A, Diallo SY, Ryzhov IO, Yilmaz L, Buckley S, Miller JA, eds. Proc. 2014 Winter Simulation Conf. (IEEE, Piscataway, NJ), 3809–3820.
- Semelhago, M., Nelson, B. L., Wächter, A., & Song, E. (2017, December). Computational methods for optimization via simulation using gaussian markov random fields. In 2017 Winter Simulation Conference (WSC) (pp. 2080-2091). IEEE.
- Mark Semelhago*, Barry L. Nelson, Eunhye Song, Andreas Waechter (2020) Rapid Optimization via Simulation with Gaussian Markov Random Fields, INFORMS Journal on Computing 33 (3) 915-930.

- Xinru Li*, Eunhye Song (2021) Projected Gaussian Markov Improvement Algorithm for High-dimensional Discrete Optimization via Simulation, Submitted.
- Salemi P, Staum J, Nelson BL (2013) Generalized integrated Brownian fields for simulation metamodeling. Proc. 2013 Winter Simulation Conf. (IEEE, Piscataway, NJ).
- P. Salemi, B. L. Nelson and J. Staum, "Generalized Integrated Brownian Fields for Simulation Metamodeling," Operations Research (2019), 874-891.
- Ding, Liang, and Xiaowei Zhang. "Sample and Computationally Efficient Simulation Metamodeling in High Dimensions." arXiv preprint arXiv:2010.06802 (2020).
- Ding, Liang, Rui Tuo, and Xiaowei Zhang. "High-Dimensional Simulation Optimization via Brownian Fields and Sparse Grids." arXiv preprint arXiv:2107.08595 (2021).

Contributions

- create a GRF that is appropriate for the DOvS problem.
- extend the GMRF approach to a multi resolution framework that can be used to solve DOvS problems with vast numbers of solutions.
- demonstrate that expected improvement (EI) combined with our framework can provide effective inference for terminating the search when the estimated optimality gap is small enough.
- Both algorithms are shown to converge to a globally optimal solution as simulation effort increases under very mild conditions (essentially finite variance).

Outline

Notations and Conditional Independence

For $C \in \mathcal{I} = \{1, \ldots, n\}$, define $\boldsymbol{x}_C = \{x_i : i \in C\}$. With $-C$ we denote the set $\mathcal{I} - C$, so that $\boldsymbol{x}_{-C} = \{x_i : i \in -C\}$. For two sets A and B, then $A \backslash B = \{i : i \in A \text{ and } i \notin B\}$.

Conditional Independence

Two variables x and y are called conditionally independent given z, iff $\pi(x, y \mid z) = \pi(x \mid z)\pi(y \mid z)$. We write this as

$$x \perp y \mid z.$$

iff $\pi(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{z}) = \pi(\boldsymbol{x} \mid \boldsymbol{z})\pi(\boldsymbol{y} \mid \boldsymbol{z})$, which we write as $\boldsymbol{x} \perp \boldsymbol{y} \mid \boldsymbol{z}$.

Factorization Criterion for Conditional Independence

$$x \perp y \mid z \iff \pi(x, y, z) = f(x, z)g(y, z)$$

for some functions f and g, and for all z with $\pi(z) > 0$.

e.g. For $\pi(x, y, z) \propto \exp(x + xz + yz)$, on some bounded region.

Undirected graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- neighbors of node i

$$\text{ne}(i) = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}$$

- neighbors of node set A

$$\text{ne}(A) = \bigcup_{i \in A} \text{ne}(i) \backslash A$$

- path from i_1 to i_m: A path from i_1 to i_m is a sequence of distinct nodes in $\mathcal{V}, i_1, i_2, \ldots, i_m$, for which $(i_j, i_{j+1}) \in \mathcal{E}$ for $j = 1, \ldots, m-1$.
- A subset $C \subset \mathcal{V}$ separates two nodes $i \notin C$ and $j \notin C$, if every path from i to j contains at least one node from C.
- Two disjoint sets $A \subset \mathcal{V} \backslash C$ and $B \subset \mathcal{V} \backslash C$ are separated by C, if all $i \in A$ and $j \in B$ are separated by C, i.e., we cannot walk
- subgraph over node set A:

$$\mathcal{E}^A = \{\{i, j\} \in \mathcal{E} \text{ and } \{i, j\} \in A \times A\}$$

The Normal Distribution

The density of a normal random variable $\boldsymbol{x} = (x_1, \ldots, x_n)^T, n < \infty$, with mean $\boldsymbol{\mu}(n \times 1$ vector) and SPD (Symmetric positive-definite) covariance matrix $\boldsymbol{\Sigma}(n \times n$ matrix), is

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad \boldsymbol{x} \in \mathbb{R}^n$$

Here, $\mu_i = \mathrm{E}(x_i), \Sigma_{ij} = \mathrm{Cov}(x_i, x_j), \Sigma_{ii} = \mathrm{Var}(x_i) > 0$ and $\mathrm{Corr}(x_i, x_j) = \Sigma_{ij}/(\Sigma_{ii}\Sigma_{jj})^{1/2}$. We write this as $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

We now divide $\boldsymbol{x}$ into two parts, $\boldsymbol{x} = \left(\boldsymbol{x}_A^T, \boldsymbol{x}_B^T\right)^T$, and split $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{pmatrix}$$

- $\boldsymbol{x}_A \sim \mathcal{N}\left(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_{AA}\right)$
- $\boldsymbol{\Sigma}_{AB} = \mathbf{0}$ iff $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are independent.
- The conditional distribution $\pi\left(\boldsymbol{x}_A \mid \boldsymbol{x}_B\right)$ is $\mathcal{N}\left(\boldsymbol{\mu}_{A|B}, \boldsymbol{\Sigma}_{A|B}\right)$, where

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\left(\boldsymbol{x}_B - \boldsymbol{\mu}_B\right) \text{ and}$$
$$\boldsymbol{\Sigma}_{A|B} = \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}$$

- If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{x}' \sim \mathcal{N}\left(\boldsymbol{\mu}', \boldsymbol{\Sigma}'\right)$ are independent, then $\boldsymbol{x} + \boldsymbol{x}' \sim \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{\mu}', \boldsymbol{\Sigma} + \boldsymbol{\Sigma}'\right)$.

GMRF: Intuition

x is a GMRF wrt $\mathcal{G}$

Let $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ have a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Define the labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, n\}$ and $\mathcal{E}$ be such that there is no edge between node i and j iff $x_i \perp x_j \mid \boldsymbol{x}_{-ij}$, where $\boldsymbol{x}_{-ij}$ is short for $\boldsymbol{x}_{-\{i,j\}}$. Then we say that $\boldsymbol{x}$ is a GMRF wrt $\mathcal{G}$.

Precision Matrix and Conditional Independence

$$\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$$

Let $\boldsymbol{x}$ be normal distributed with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} > 0$. Then for $i \neq j$,

$$x_i \perp x_j \mid \boldsymbol{x}_{-ij} \iff Q_{ij} = 0.$$

GMRF: Definition

A random vector $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ is called a GMRF wrt a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} > 0$, iff its density has the form

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2} |\boldsymbol{Q}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

and

$$Q_{ij} \neq 0 \quad \Longleftrightarrow \quad \{i, j\} \in \mathcal{E} \quad \text{for all} \quad i \neq j$$

Any normal distribution with SPD covariance matrix is also a GMRF and vice versa.

GMRF: Inference

Let $\boldsymbol{x}$ be a GMRF wrt $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} > 0$, then

$$E\left(x_i \mid \boldsymbol{x}_{-i}\right) = \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}\left(x_j - \mu_j\right),$$

$$\operatorname{Prec}\left(x_i \mid \boldsymbol{x}_{-i}\right) = Q_{ii} \text{ and}$$

$$\operatorname{Corr}\left(x_i, x_j \mid \boldsymbol{x}_{-ij}\right) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \quad i \neq j.$$

Markov properties of GMRFs

Theorem 2.4 Let $\boldsymbol{x}$ be a GMRF wrt $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Then the following are equivalent. The pairwise Markov property:

$$x_i \perp x_j \mid \boldsymbol{x}_{-ij} \quad \text{if } \{i,j\} \notin \mathcal{E} \text{ and } i \neq j$$

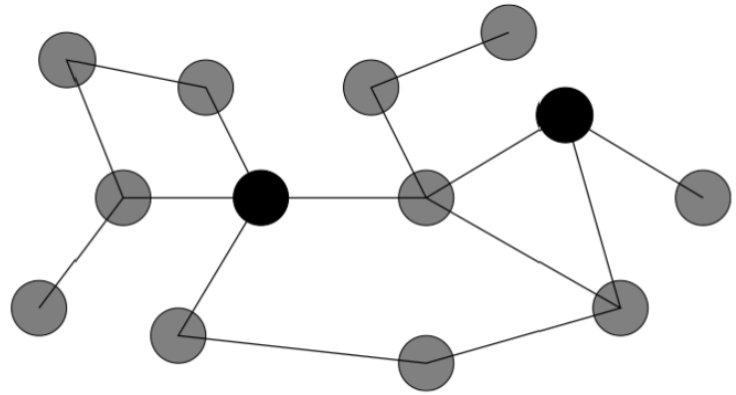
The local Markov property:

$$x_i \perp \boldsymbol{x}_{-\{i,ne(i)\}} \mid \boldsymbol{x}_{ne(i)} \quad \text{for every } i \in \mathcal{V}$$

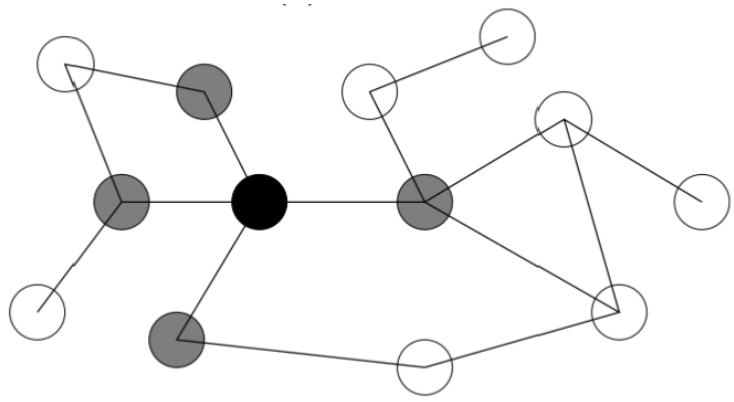
The global Markov property:

$$\boldsymbol{x}_A \perp \boldsymbol{x}_B \mid \boldsymbol{x}_C$$

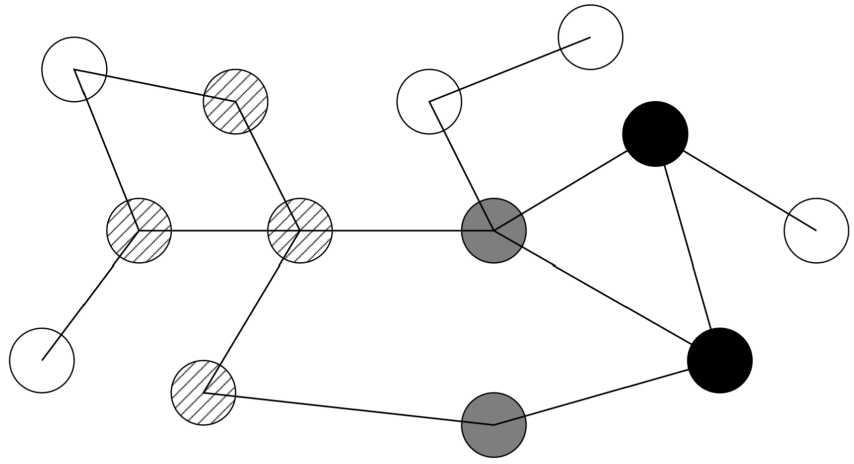
for all disjoint sets A, B and C where C separates A and B, and A and B are non-empty.



(*a*) Pairwise

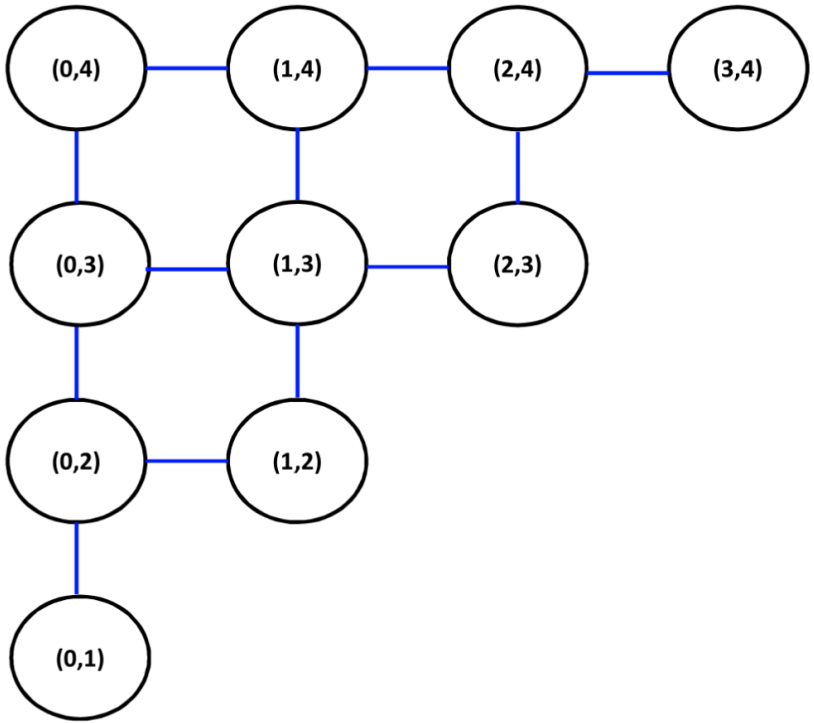


(*b*) Local



(*c*) Global

Local Markov Property on Lattice



The Markov property does not imply that nodes far away from one another are independent, but rather that if we know the value of the GMRF at nodes close by, then we can ignore nodes farther away conditional on those values.

Conditional properties of GMRFs

We split the indices into the nonempty sets A and denote by B the set $-A$, so that

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_A \\ \boldsymbol{x}_B \end{pmatrix}$$

Partition the mean and precision accordingly,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \quad \text{and} \quad \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{AA} & \boldsymbol{Q}_{AB} \\ \boldsymbol{Q}_{BA} & \boldsymbol{Q}_{BB} \end{pmatrix}$$

Let $\boldsymbol{x}$ be a GMRF wrt $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} > 0$. Let $A \subset \mathcal{V}$ and $B = \mathcal{V} \backslash A$ where $A, B \neq \emptyset$. The conditional distribution of $\boldsymbol{x}_A \mid \boldsymbol{x}_B$ is then a GMRF wrt the subgraph $\mathcal{G}^A$ with mean $\boldsymbol{\mu}_{A|B}$ and precision matrix $\boldsymbol{Q}_{A|B} > 0$, where

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A - \boldsymbol{Q}_{AA}^{-1} \boldsymbol{Q}_{AB} \left(\boldsymbol{x}_B - \boldsymbol{\mu}_B \right)$$

and

$$\boldsymbol{Q}_{A|B} = \boldsymbol{Q}_{AA}.$$

This is a powerful result for two reasons.

- First, we have explicit knowledge of $\boldsymbol{Q}_{A|B}$ through the principal matrix $\boldsymbol{Q}_{AA}$, so no computation is needed to obtain the conditional precision matrix. Constructing the subgraph $\mathcal{G}^A$ does not change the structure; it just removes all nodes not in A and the corresponding edges.
- Secondly, since Q_{ij} is zero unless $j \in \text{ne}(i)$, the conditional mean only depends on values of $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ in $A \cup \text{ne}(A)$. This is a great advantage if A is a small subset of $\mathcal{V}$.

The canonical parameterization

Canonical parameterization

A GMRF x wrt $\mathcal{G}$ with canonical parameters $\boldsymbol{b}$ and $\boldsymbol{Q} > 0$ has density

$$\pi(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\right),$$

i.e., the precision matrix is $\boldsymbol{Q}$ and the mean is $\boldsymbol{\mu} = \boldsymbol{Q}^{-1}\boldsymbol{b}$. We write the canonical parameterization as $\boldsymbol{x} \sim \mathcal{N}_C(\boldsymbol{b}, \boldsymbol{Q})$. The relation to the normal is that $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}\right) = \mathcal{N}_C(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q})$.

Let $\boldsymbol{x} \sim \mathcal{N}_C(\boldsymbol{b}, \boldsymbol{Q})$, then

$$\boldsymbol{x}_A \mid \boldsymbol{x}_B \sim \mathcal{N}_C\left(\boldsymbol{b}_A - \boldsymbol{Q}_{AB}\boldsymbol{x}_B, \boldsymbol{Q}_{AA}\right).$$

The result is useful for computing conditional densities.

GMRF Construction for DOvS Problem

Construction of the graph:

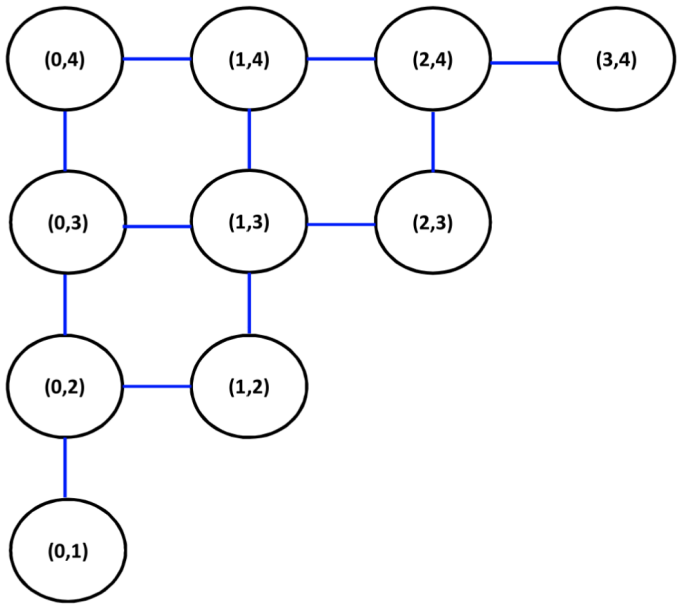
- defining the nodes of the graph $\mathscr{V}$ to be $\mathscr{X}$
- defining the edges = construction of neighborhoods

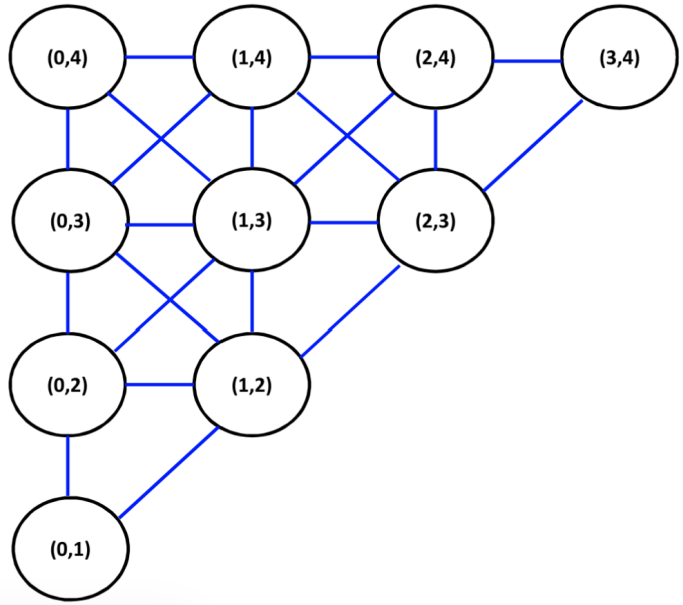
$$\mathcal{N}(\mathbf{x}) = \{\mathbf{x}' \in \mathscr{X} : \|\mathbf{x} - \mathbf{x}'\|_2 = 1\}$$

up to $2d$ neighbors in d dimensions. For this neighborhood, the fraction of nonzero entries in the precision matrix $\mathbf{Q}$ is bounded above by $(2d+1)/|\mathscr{X}|$, which is very small for large problems.

$$\mathcal{N}(\mathbf{x}) = \{\mathbf{x}' \in \mathscr{Z} : \|\mathbf{x} - \mathbf{x}'\|_2 = 1\}$$

up to $3^d - 1$ neighbors in d dimensions.





Because we are particularly interested in DOvS problems with large feasible regions and because (roughly speaking) the more sparse $\mathbf{Q}$ is, the larger the DOvS problems we can solve, our algorithm will use the neighborhood structure $\mathcal{N}(\mathbf{x})$.

Construction of Q

Recall the definition of GMRF,

$$Q_{ij} \neq 0 \quad \Longleftrightarrow \quad \{i,j\} \in \mathcal{E} \quad \text{for all} \quad i \neq j.$$

A standard approach to specify $\mathbf{Q}$ is to have the entries of $\mathbf{Q}$ given by a function $p\left(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}\right)$, where θ is a vector of parameters; that is, $\mathrm{Q}_{ij} \triangleq p\left(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}\right)$. For the neighborhood $\mathcal{N}(\mathbf{x})$, we propose $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \ldots, \theta_d)^\top$ and

$$p\left(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}\right) = \begin{cases} \theta_0, & \text{if } \mathbf{x} = \mathbf{x}' \\ -\theta_0\theta_j, & \text{if } |\mathbf{x} - \mathbf{x}'| = \mathbf{e}_j \\ 0, & \text{otherwise} \end{cases}$$

for $\mathbf{x}, \mathbf{x}' \in \mathbb{Z}^d$, where $\mathbf{e}_j$ is the j th standard basis vector.

Properties of the constructed Q

Recall that

$$\text{Prec}\left(x_i \mid \boldsymbol{x}_{-i}\right) = Q_{ii} \text{ and}$$

$$\text{Corr}\left(x_i, x_j \mid \boldsymbol{x}_{-ij}\right) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \quad i \neq j.$$

Thus, θ_0 is the conditional precision of each solution, and θ_j is the conditional correlation between solutions that differ by one in the j th coordinate direction.

That the conditional correlations can depend on the coordinate direction is important to allow for response surfaces that change more rapidly in one direction as compared with another.

Restrictions on Parameterization

the conditional precisions must be positive, it follows that

$$\theta_0 > 0$$

neighbors have nonnegative conditional correlations

$$\theta_1, \theta_2, \ldots, \theta_d \geq 0$$

conditional correlations must be less than one

$$\theta_j \leq 1 \text{ for } j = 1, 2, \ldots, d$$

Finally, $\mathbf{Q}$ should be positive-definite.

Now, $\mathbf{Q}$ is a nonsingular M matrix, so its inverse is nonnegative; that is, $\left[\mathbf{Q}^{-1}\right]_{ij} \geq 0$ for all i and j. In other words, **there are no negative unconditional correlations among nodes in the GMRF.**

Modeling the Mean

$$\text{minimize } y(\mathbf{x}) \triangleq \mathrm{E}[Y(\mathbf{x})] \text{ subject to } \mathbf{x} \in \mathscr{X}$$

$$Y_j(\mathbf{x}) = y(\mathbf{x}) + \epsilon_j(\mathbf{x})$$

where $\{\epsilon_j(\mathbf{x})\}$ are independent and identically distributed (i.i.d.) with mean zero and finite variance that may depend on $\mathbf{x}$. In the following we also assume that the $\epsilon_j(\mathbf{x})$ are normally distributed.

Bayesian Prior Model of the Means

Let $\mathbf{y}$ denote the vector of objective function values $(y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_n))^\top$. Of course, $\mathbf{y}$ is unknown, so we model it as a realization of the GMRF

$$\mathbb{Y} \triangleq (\mathbb{Y}(\mathbf{x}_1), \mathbb{Y}(\mathbf{x}_2), \ldots, \mathbb{Y}(\mathbf{x}_n))^\top \sim \mathcal{N}\left(\mu, \mathbf{Q}(\boldsymbol{\theta})^{-1}\right)$$

More Notations

- "design point": refer to a feasible solution that has been simulated
- "point" = "feasible solution"
- $\Xi_2 \subseteq \mathscr{X}$ denote the current set of design points
- $\Xi_1 = \mathscr{X} \backslash \Xi_2$ is the set of feasible solutions that have not been simulated
- use "1" as a subscript to denote quantities associated with the set Ξ_1 and " 2 " as a subscript to denote quantities associated with the set Ξ_2.
- $n_1 = |\Xi_1|$ and $n_2 = |\Xi_2|$

Intermediate Inference

(In the running process) partition the vectors $\mathbf{y}, \mathbb{Y}, \mu$ and the precision matrix $\mathbf{Q}(\boldsymbol{\theta})$ as

$$\begin{pmatrix} \mathbb{Y}_1 \\ \mathbb{Y}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \mathbf{Q}_{11}(\boldsymbol{\theta}) & \mathbf{Q}_{12}(\boldsymbol{\theta}) \\ \mathbf{Q}_{12}(\boldsymbol{\theta})^\top & \mathbf{Q}_{22}(\boldsymbol{\theta}) \end{pmatrix}^{-1} \right)$$

Let $\overline{\mathscr{Y}}_2$ be the vector of sample means of the simulation output at the design points and represent $\overline{\mathscr{Y}}_2$ as a realization of the GMRF

$$\mathbb{Y}_2^\epsilon = \mathbb{Y}_2 + \epsilon$$

with $\mathbb{Y}_2$ and ϵ independent and $\epsilon \sim \mathcal{N}\left(\overrightarrow{\mathbf{0}}_{n_2 \times 1}, \mathbf{Q}_\epsilon^{-1}\right)$, where $\mathbf{Q}_\epsilon$ is the intrinsic precision matrix of the noise inherent to the stochastic simulation output $\overline{\mathscr{Y}}_2$.

$\mathbf{Q}$ & $\mathbf{Q}_\epsilon$ Estimation (MLE)

The values in $\mathbf{Q}_\epsilon$ also depend on how many replications have been averaged, which need not be the same at all design points.

Ignored.

Electronic Companion to Gaussian Markov Random Fields for Discrete Optimization via Simulation: Theory and Computation

EC.1. Conditional Distribution Proof

In this section we prove Theorem 1, which establishes the conditional distribution of $(\mathbb{Y}_1, \mathbb{Y}_2)$ given observed $\mathbb{Y}_2^\epsilon$ in (4). We do this by first deriving the joint distribution of $(\mathbb{Y}_1, \mathbb{Y}_2, \mathbb{Y}_2^\epsilon)$ and then applying Lemma 2.1 of Rue and Held (2005) to obtain the conditional distribution. To simplify the notation we suppress the dependence of $\mathbf{Q}$ on $\boldsymbol{\theta}$.

Notice that $\mathbb{Y}_1$ and $\mathbb{Y}_2^\epsilon$ are conditionally independent, given $\mathbb{Y}_2$, because $\mathbb{Y}$ is a GMRF and is also independent of the intrinsic noise. From Theorem 2.5 in Rue and Held (2005), the conditional distribution of $\mathbb{Y}_1$ given $\mathbb{Y}_2 = \mathbf{y}_2$ is

$$\mathcal{N}\left(\boldsymbol{\mu}_1 - \mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}(\mathbf{y}_2 - \boldsymbol{\mu}_2), \mathbf{Q}_{11}^{-1}\right).$$

From our assumption about the simulation output process, the conditional distribution of $\mathbb{Y}_2^\epsilon$ given $\mathbb{Y}_2 = \mathbf{y}_2$ is $\mathcal{N}(\mathbf{y}_2, \mathbf{Q}_\epsilon^{-1})$. Further,

$$\mathbb{Y}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) = \mathcal{N}\left(\boldsymbol{\mu}_2, [\mathbf{Q}_{22} - \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}]^{-1}\right)$$

using standard results for the inverse of a partitioned matrix. Therefore, the joint distribution $f(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_2^\epsilon)$ satisfies

$$\begin{aligned} f(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_2^\epsilon) \propto \exp&\left\{-\frac{1}{2}\left[\mathbf{y}_1 - (\boldsymbol{\mu}_1 - \mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}(\mathbf{y}_2 - \boldsymbol{\mu}_2))\right]^\top \mathbf{Q}_{11}\left[\mathbf{y}_1 - (\boldsymbol{\mu}_1 - \mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}(\mathbf{y}_2 - \boldsymbol{\mu}_2))\right]\right\} \\ &\times \exp\left\{-\frac{1}{2}(\mathbf{y}_2 - \boldsymbol{\mu}_2)^\top (\mathbf{Q}_{22} - \mathbf{Q}_{21}\mathbf{Q}_{11}^{-1}\mathbf{Q}_{12})(\mathbf{y}_2 - \boldsymbol{\mu}_2)\right\} \\ &\times \exp\left\{-\frac{1}{2}(\mathbf{y}_2^\epsilon - \mathbf{y}_2)^\top \mathbf{Q}_\epsilon (\mathbf{y}_2^\epsilon - \mathbf{y}_2)\right\}. \end{aligned} \tag{EC.1}$$

A significant quantity of matrix algebra leads to

$$\begin{aligned} f(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_2^\epsilon) \propto \exp&\left\{-\frac{1}{2}\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \mathbf{y}_2^\epsilon\end{pmatrix}^\top \begin{pmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12} & \mathbf{0}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}+\mathbf{Q}_\epsilon & -\mathbf{Q}_\epsilon\\ \mathbf{0} & -\mathbf{Q}_\epsilon & \mathbf{Q}_\epsilon\end{pmatrix}\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \mathbf{y}_2^\epsilon\end{pmatrix}\right. \\ &\left. + \begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \mathbf{y}_2^\epsilon\end{pmatrix}^\top \begin{pmatrix}\mathbf{Q}_{11}\boldsymbol{\mu}_1 + \mathbf{Q}_{12}\boldsymbol{\mu}_2\\ \mathbf{Q}_{21}\boldsymbol{\mu}_1 + \mathbf{Q}_{22}\boldsymbol{\mu}_2\\ \mathbf{0}\end{pmatrix}\right\}. \end{aligned} \tag{EC.2}$$

Rue and Held (2005) refer to (EC.2) as being in canonical form, for which their Lemma 2.1 can be used to show that the conditional distribution of $(\mathbb{Y}_1, \mathbb{Y}_2)$ given $\mathbb{Y}_2^\epsilon$ is as in (4).

Inference

Theorem 1. The conditional distribution of $\mathbb{Y}$ given $\mathbb{Y}_2^\epsilon = \overline{\mathscr{Y}}_2$ is

$$\mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \overline{\mathbf{Q}}(\boldsymbol{\theta})^{-1} \begin{pmatrix} \overrightarrow{\mathbf{0}}_{n_1 \times 1} \\ \mathbf{Q}_\epsilon \left(\overline{\mathscr{Y}}_2 - \mu_2\right) \end{pmatrix}, \overline{\mathbf{Q}}(\boldsymbol{\theta})^{-1}\right)$$

where

$$\overline{\mathbf{Q}}(\boldsymbol{\theta}) \triangleq \begin{pmatrix} \mathbf{Q}_{11}(\boldsymbol{\theta}) & \mathbf{Q}_{12}(\boldsymbol{\theta}) \\ \mathbf{Q}_{12}(\boldsymbol{\theta})^\top & \mathbf{Q}_{22}(\boldsymbol{\theta}) \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{n_1 \times n_1} & \mathbf{0}_{n_1 \times n_2} \\ \mathbf{0}_{n_1 \times n_2}^\top & \mathbf{Q}_\epsilon \end{pmatrix}$$

is the conditional precision matrix and $0_{n_i \times n_j}$ is the $n_i \times n_j$ matrix of zeros.
The sparsity of $\overline{\mathbf{Q}}(\boldsymbol{\theta})$ is inherited from the sparsity of $\mathbf{Q}(\boldsymbol{\theta})$ and $\mathbf{Q}_\epsilon$.

The Defined EI Criterion

Let $\mathbf{x}^\star$ denote the unknown best of the n feasible solutions, which we assume for simplicity of exposition is unique: $\mathbf{x}^\star = \arg\min_{\mathbf{x}} y(\mathbf{x})$. Let $\tilde{\mathbf{x}}$ be the selected solution by whatever method.

EI for a feasible solution x

$$\mathrm{EI}^t(\mathbf{x}) = \mathrm{E}\left[\max\left\{0, \mathbb{Y}\left(\tilde{\mathbf{x}}^t\right) - \mathbb{Y}(\mathbf{x})\right\} \mid \overline{\mathscr{Y}}_2^t\right]$$

where we have appended a superscript t to indicate quantities available at the end of the t th iteration of an algorithm.

$\tilde{\mathbf{x}}^t$ can be the sample best solution (smallest sample mean) through iteration t.

Closed Form of "complete expected improvement"

Recall that the conditional distribution of $\mathbb{Y}$ given $\mathbb{Y}_2^{\epsilon} = \overline{\mathscr{Y}}_2$ is

$$\mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \overline{\mathbf{Q}}(\boldsymbol{\theta})^{-1} \begin{pmatrix} \overrightarrow{\mathbf{0}}_{n_1 \times 1} \\ \mathbf{Q}_{\epsilon}\left(\overline{\mathscr{Y}}_2 - \mu_2\right) \end{pmatrix}, \overline{\mathbf{Q}}(\boldsymbol{\theta})^{-1}\right).$$

Given the simulation output $\overline{\mathscr{Y}}_2^t$, the conditional joint distribution of $\mathbb{Y}\left(\tilde{\mathbf{x}}^t\right)$ and $\mathbb{Y}(\mathbf{x})$ is bivariate normal.

- Denote the conditional means by $M^t\left(\tilde{\mathbf{x}}^t\right)$ and $M^t(\mathbf{x})$
- the conditional variances by $V^t\left(\tilde{\mathbf{x}}^t\right)$ and $V^t(\mathbf{x})$, and the conditional correlation by $\rho^t\left(\tilde{\mathbf{x}}^t, \mathbf{x}\right)$.
- Let

$$V^t\left(\tilde{\mathbf{x}}^t, \mathbf{x}\right) \triangleq V^t\left(\tilde{\mathbf{x}}^t\right) + V^t(\mathbf{x}) - 2\rho^t\left(\tilde{\mathbf{x}}^t, \mathbf{x}\right)\sqrt{V^t\left(\tilde{\mathbf{x}}^t\right) V^t(\mathbf{x})}$$

be the conditional variance of the difference $\mathbb{Y}\left(\tilde{\mathbf{x}}^t\right) - \mathbb{Y}(\mathbf{x})$.

Then the CEI of solution $\mathbf{x}, \mathrm{CEI}(\mathbf{x})$, is

$$\mathrm{CEI}^t(\mathbf{x}) = \left(M^t\left(\tilde{\mathbf{x}}^t\right) - M^t(\mathbf{x})\right)\Phi\left(\frac{M^t\left(\tilde{\mathbf{x}}^t\right) - M^t(\mathbf{x})}{\sqrt{V^t\left(\tilde{\mathbf{x}}^t, \mathbf{x}\right)}}\right) + \sqrt{V^t\left(\tilde{\mathbf{x}}^t, \mathbf{x}\right)}\phi\left(\frac{M^t\left(\tilde{\mathbf{x}}^t\right) - M^t(\mathbf{x})}{\sqrt{V^t\left(\tilde{\mathbf{x}}^t, \mathbf{x}\right)}}\right).$$

Detailed discussions on the EI are ignored here.

Stopping Criterion

$$\max_{\mathbf{x} \neq \tilde{\mathbf{x}}^t} CEI^t(\mathbf{x}) \leq \delta$$

Comparison with PCS and PGS:

$$\text{(PCS)}\, \Pr\left\{ \tilde{\mathbf{x}} = \mathbf{x}^\star \mid \min_{\mathbf{x} \neq \mathbf{x}^\star} y(\mathbf{x}) - y\left(\mathbf{x}^\star\right) \geq \delta \right\}$$

$$\text{(PGS)}\, \Pr\left\{ y(\tilde{\mathbf{x}}) - y\left(\mathbf{x}^\star\right) \leq \delta \right\} \geq 1 - \alpha$$

The proposed approach is more akin to PGS than PCS in that we make no assumption about the gap between the best and next-best feasible solution, but we do require the user to provide a smallest practically significant difference δ measured in the same units as the response $y(\mathbf{x})$.

Outline

Solution-level Gaussian Markov improvement algorithm

- 0. Generate a set of k_s design points. Simulate r replications for each design point and use the simulation output to calculate the MLE estimates.
- 1. Let $\tilde{\mathbf{x}}$, the current sample-best solution, be the design point with the smallest sample mean.
- 2. Calculate the CEI with respect to $\tilde{\mathbf{x}}$ for each candidate feasible solution. If $\max_{x \neq \tilde{x}} \text{CEI}(\mathbf{x}) \leq \delta$, then go to step 5. Otherwise, go to step 3.
- 3. Simulate r replications at both $\tilde{\mathbf{x}}$ and the candidate feasible solution $\mathbf{x}^*_{\text{CEI}}$ that maximize the CEI over the set of all candidate feasible solutions.

- 4. Update the simulation output at $\tilde{\mathbf{x}}$ with the new replications. If $\mathbf{x}^*_{\text{CEI}}$ is a design point, then update the simulation output at $\mathbf{x}^*_{\text{CEI}}$ and go to step 1. If $\mathbf{x}^*_{\text{CEI}}$ is not a design point, then add $\mathbf{x}^*_{\text{CEI}}$ to the set of design points, add the simulation output obtained at $\mathbf{x}^*_{\text{CEI}}$ to the collection of simulation output, and go to step 1.
- 5. Return $\tilde{\mathbf{x}}$ as the estimated optimal solution.

The MLE estimates are not updated dynamically.

Estimated Conditional Distribution

$$\mathcal{N}\left(\hat{\beta}_0\mathbf{1}_{n\times 1} + \overline{\mathbf{Q}}(\hat{\theta})^{-1}\begin{pmatrix} \overrightarrow{\mathbf{0}}_{n_1\times 1} \\ \hat{\mathbf{Q}}_\epsilon\left(\overline{\mathscr{Y}}_2 - \hat{\beta}_0\mathbf{1}_{n_2\times 1}\right)\end{pmatrix}, \overline{\mathbf{Q}}(\hat{\theta})^{-1}\right),$$

where

$$\overline{\mathbf{Q}}(\hat{\boldsymbol{\theta}}) \triangleq \begin{pmatrix} \mathbf{Q}_{11}(\hat{\boldsymbol{\theta}}) & \mathbf{Q}_{12}(\hat{\boldsymbol{\theta}}) \\ \mathbf{Q}_{12}(\hat{\boldsymbol{\theta}})^\top & \mathbf{Q}_{22}(\hat{\boldsymbol{\theta}}) \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{n_1\times n_1} & \mathbf{0}_{n_1\times n_2} \\ \mathbf{0}_{n_1\times n_2}^\top & \hat{\mathbf{Q}}_\epsilon \end{pmatrix}$$

Computation of $\overline{\mathbf{Q}}(\hat{\theta})^{-1}$

The matrix is sparse, and with our proposed neighborhood structure, the fraction of nonzero elements is bounded above by $(2d+1)/n$. To compute all $n-1$ CEIs we need only the n diagonal elements of $\overline{\mathbf{Q}}^{-1}(\hat{\boldsymbol{\theta}})$ and the $n-1$ off-diagonal elements of one column.

Asymptotic Convergence

Assumptions

1. $y(\mathbf{x}) > -\infty$ for all $\mathbf{x} \in \mathscr{X}$
2. $0 < \text{Var}[Y(\mathbf{x})] < +\infty$ for all $\mathbf{x} \in \mathscr{X}$
3. The initial estimated precision matrix $\mathbf{Q}(\hat{\boldsymbol{\theta}})$ is positive-definite and is not updated thereafter.

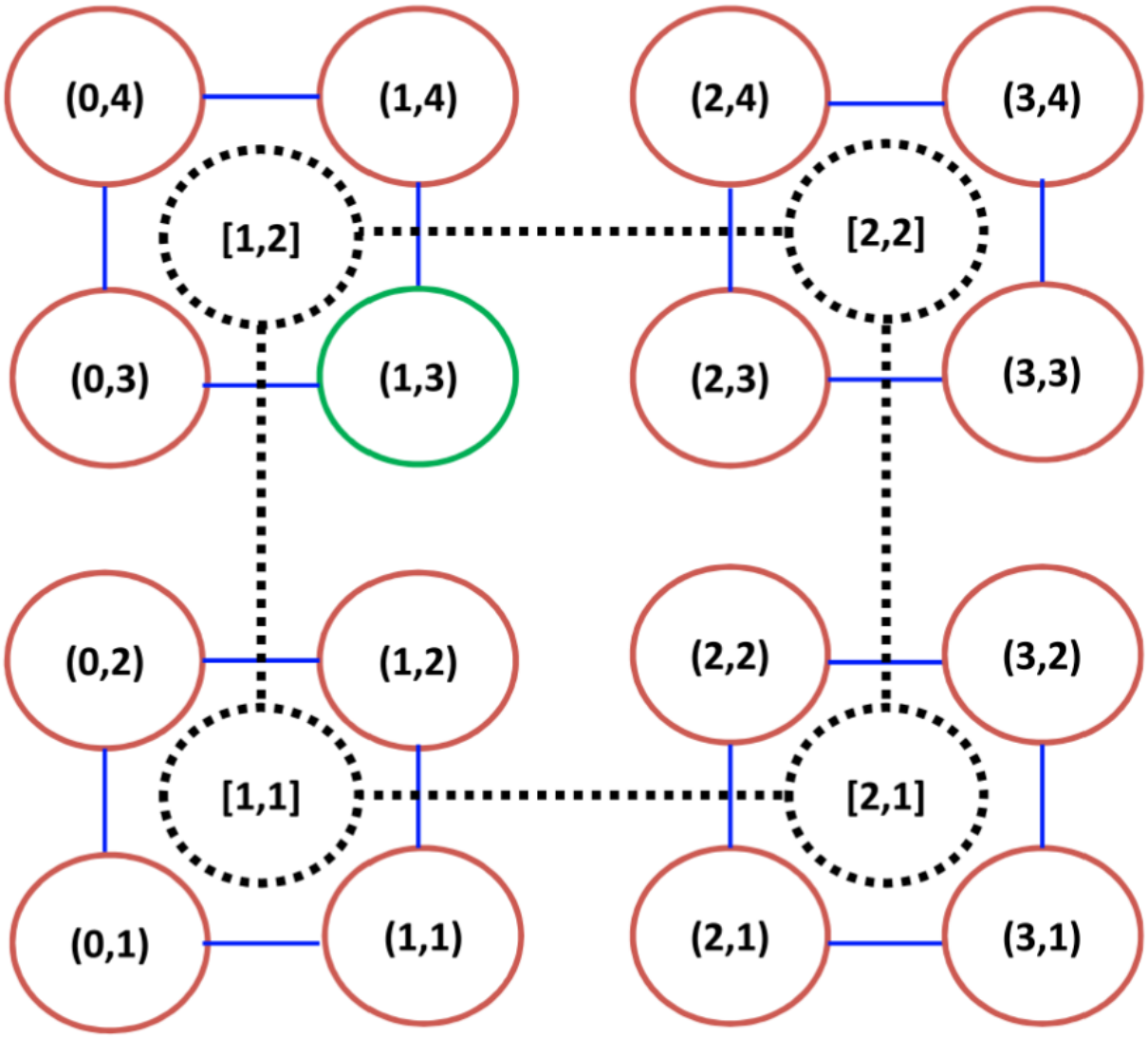
Theorem 2.

The GMIA algorithm without a stopping condition simulates each solution $\mathbf{x} \in \mathbb{X}$ infinitely often with probability one as the number of iterations goes to infinity.

Theorem 2 does not depend on our chosen neighborhood structure being "correct," nor does it depend on having "good" parameter estimates for the GMRF.

Outline

Solution Level v.s. Region Level



The Multi-resolution Framework and Algorithm

- First partition $\mathscr{X}$ into m disjoint regions $\mathscr{P}_1, \mathscr{P}_2, \ldots, \mathscr{P}_m$
- Nodes of a region-level GMRF represent a measure of the overall solution quality within each region with a neighborhood structure again defined by proximity (i.e., adjacent regions).
- The region-level GMRF provides global guidance by facilitating a CEI comparison among regions.
- The quality of individual solutions within a region $\mathscr{P}_j$ is represented by a solution-level GMRF except that the GMRs for $\mathscr{P}_j$ and $\mathscr{P}_\ell$ for $j \neq \ell$ are assumed to be independent.

Region-Level Prior Model

The response associated with the region $\mathscr{P}_\ell$ is

$$z\left(\mathscr{P}_\ell\right) = \left|\mathscr{P}_\ell\right|^{-1} \sum_{\mathbf{x} \in \mathscr{P}_\ell} y(\mathbf{x})$$

model $\mathbf{z} = \left(z\left(\mathscr{P}_1\right), z\left(\mathscr{P}_2\right), \ldots, z\left(\mathscr{P}_m\right)\right)^\top$ as as a realization of a region-level GMRF

$$\mathbb{Z} \triangleq \left(\mathbb{Z}\left(\mathscr{P}_1\right), \mathbb{Z}\left(\mathscr{P}_2\right), \ldots, \mathbb{Z}\left(\mathscr{P}_m\right)\right)^\top \sim \mathcal{N}\left(\boldsymbol{\eta}, \mathbf{T}(\boldsymbol{\tau})^{-1}\right).$$

For the solution-level GMRFs, we may choose to fit a different set of parameters for each region- $\boldsymbol{\mu}_\ell$ and $\boldsymbol{\theta}_\ell$ for $\mathscr{P}_\ell$-or assume all regions share the same solution-level parameters $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$.

MR-GMIA: Procedure

- On each iteration of the algorithm, select three regions in which to run simulations:
  - (i) the design region $\mathscr{P}_{\min}$ that contains the solution with the smallest sample mean among all simulated solutions in $\mathscr{X}$,
  - (ii) the design region $\tilde{\mathscr{P}}$ that has the smallest region-level estimator $\overline{\mathscr{L}}(\mathscr{P}_\ell)$ among $\mathscr{P}_\ell \in \Pi_2$,
  - (iii) the region $\mathscr{P}^\star_{\text{CEI}}$ that has the largest region-level CEI among all regions in $\mathscr{P}$.
- For $\mathscr{P}_{\min}$ and $\tilde{\mathscr{P}}$, and for $\mathscr{P}^\star_{\text{CEI}}$, apply the same method as in the solution-level GMIA algorithm:
  at each iteration, simulate the solution with the smallest sample mean and the solution with the largest solution-level CEI in that region.

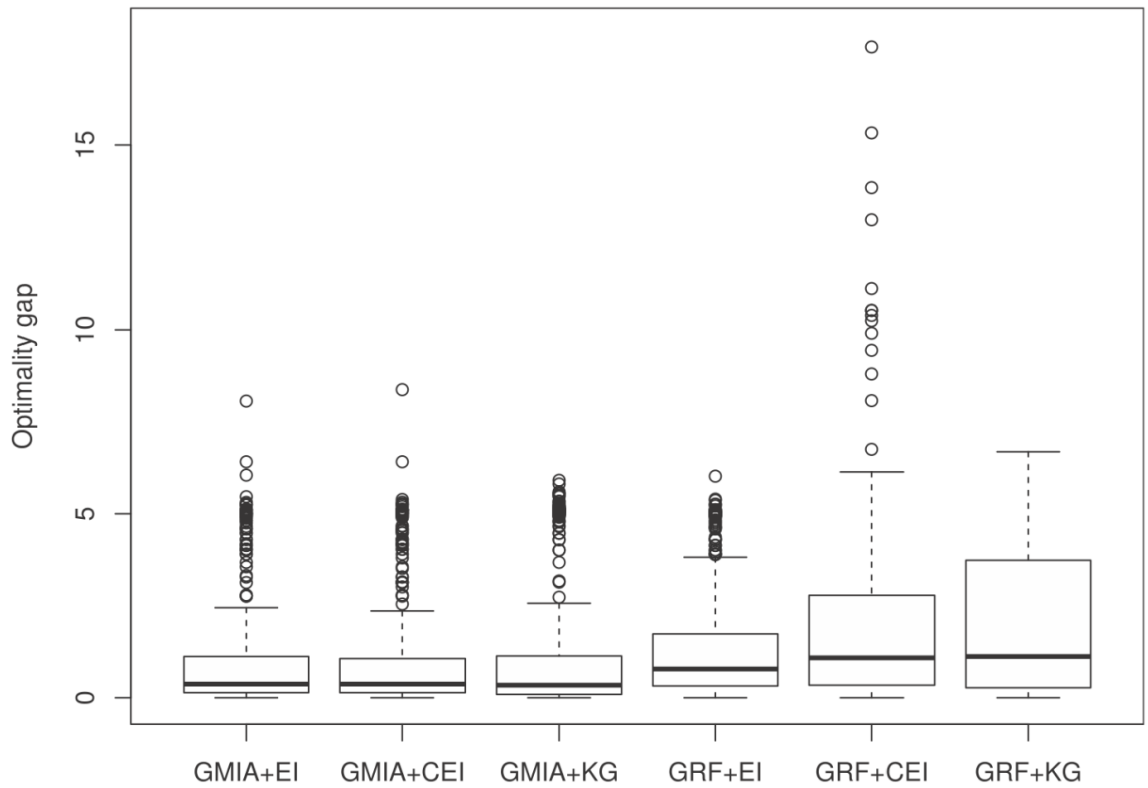
The algorithm stops when both the largest region-level CEI and the solution-level CEI fall below δ.

Outline

Experiments

- (i) GMIA combined with expected improvement provides valid inference on the remaining optimality gap at termination while achieving good finite-budget performance for a DOvS problem,
- (ii) MR-GMIA can solve a large-scale DOvS problem efficiently and effectively.
- (s, S) Inventory Problem (2d)
- Griewank Function (2d)
- Inverted Multivariate Normal Density Function (15d)

(s, S) Inventory Problem with 100×100 lattice



	GMIA+EI	GMIA+CEI	GMIA+KG	GRF+EI	GRF+CEI	GRF+KG
Mean	1.124	1.090	1.257	1.341	2.627	2.034
Std. err.	0.081	0.078	0.092	0.075	0.412	0.100

Figure 1: optimality gaps from 400 runs of all six algorithms stopped after 100 iterations

(s, S) Inventory Problem with 100×100 lattice

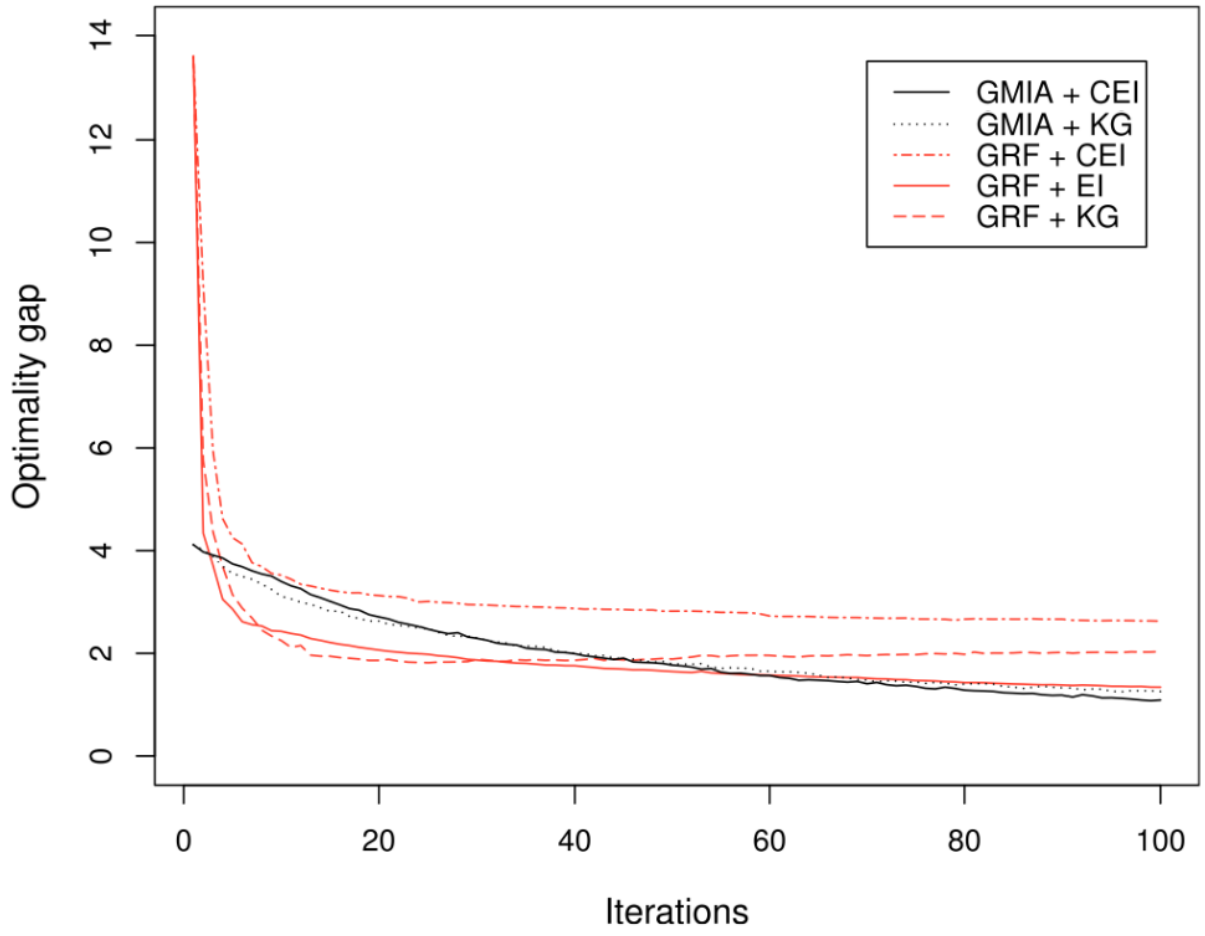


Figure 2: convergence plot of all six algorithms stopped after 100 iterations

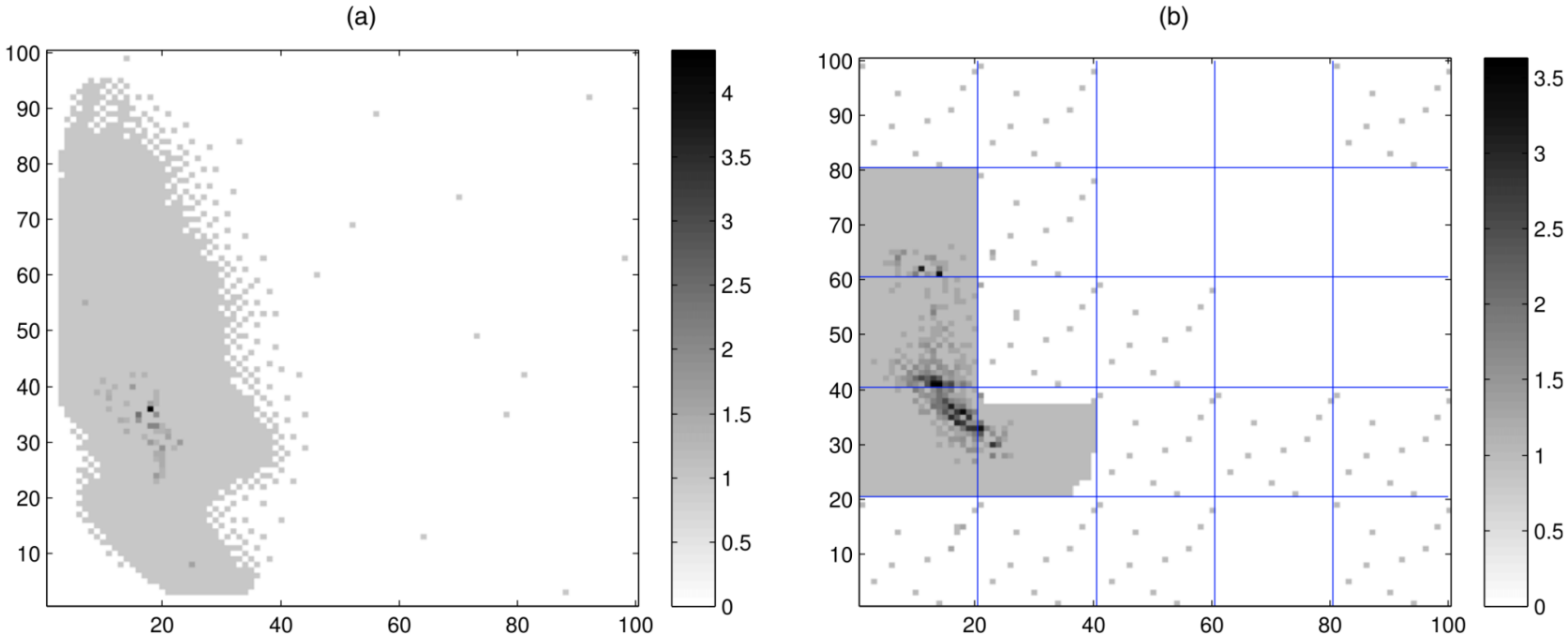
Achieving Optimality gap $\delta = 1$

Table 1. Average and Maximum of Optimality Gaps at Termination of 50 Runs of GMIA + EI/CEI, GRF + EI/CEI, and KN Given $\delta = \$1$

	GMIA + EI	GMIA + CEI	GRF + EI	GRF + CEI	KN
Average optimality gap	0.089 (0.009)	0.096 (0.012)	3.283 (0.346)	3.401 (0.343)	0.097 (0.009)
Maximum optimality gap	0.271	0.348	8.545	8.545	0.271
Average number of sampled solutions	2,775 (82)	2,750 (76)	20.9 (0.2)	20.3 (0.1)	10,000 —
Average number of replications	55,314 (1,633)	54,854 (1,526)	218 (4)	207 (2)	108,111 (185)

Notes. The average number of sampled solutions and the average number of replications spent by each algorithm until termination are also presented. The standard errors are presented in parentheses.

Comparison of GMIA and MR-GMIA.



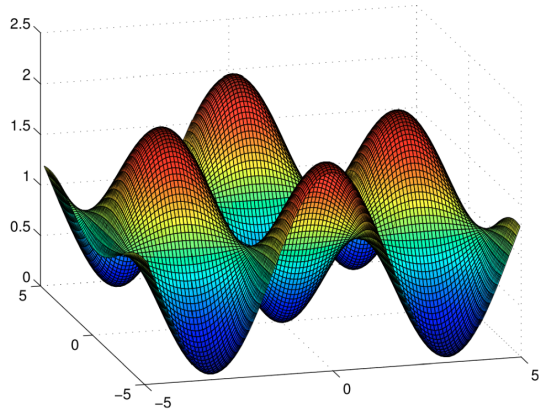
Notes. (a) GMIA. (b) MR-GMIA: Regions are numbered 1–25 bottom to top from the lower left corner.

Table 2. Average and Maximum of Optimality Gaps at Termination of 50 Runs of MR-GMIA + EI/CEI

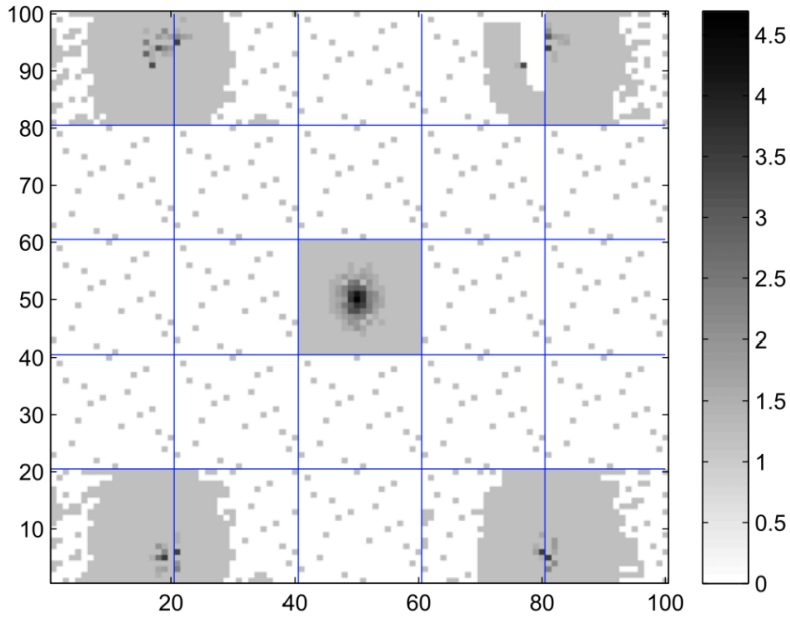
	MR-GMIA + EI	MR-GMIA + CEI
Average optimality gap	0.505 (0.065)	0.321 (0.054)
Maximum optimality gap	1.886	2.218
Average number of sampled solutions	303 (36)	525 (73)
Average number of replications	6,071 (901)	13,214 (2,188)

Notes. The average number of sampled solutions and the average number of replications spent by each algorithm until termination are also presented. The standard errors are presented in parentheses.

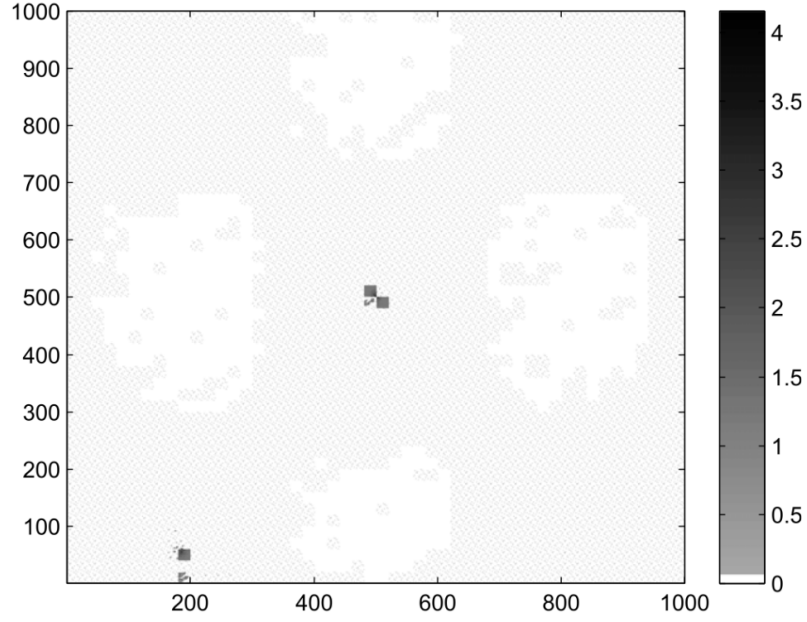
Griewank Function (2d)



(a)



(b)



Notes. (a) 100 × 100 lattice. (b) 1,000 × 1,000 lattice.

Inverted Multivariate Normal Density Function

- integer lattice with $4^{15} = 1,073,741,824$ solutions.
- The response function for this test case is a 15-dimensional inverted Gaussian density function (that does not have any local minima),

$$f(x_1, x_2, \ldots, x_d) = -\beta \exp\left\{-\gamma \sum_{j=1}^{d} j x_j^2\right\}$$

$$\mathbf{x}^\top = (x_1, x_2, \ldots, x_d).$$

- a three-resolution GMRF consisting of super region–level, region-level, and solution-level GMRFs.
- The average optimality gap of 200 runs after 10,000 iterations is 1.27 (standard error = 0.12), and
- the average number of solutions that the algorithm visited is $73,344(1,046)$, which is only 0.0068% of the feasible solution space.

Outline