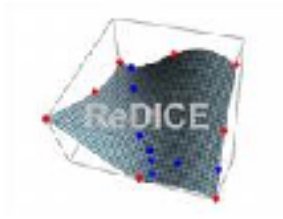
High-dimensional Bayesian multi-objective optimization with random embeddings

Mickaël Binois

Mines Saint-Étienne
Renault
ReDICE consortium

Joint work with David Ginsbourger (UniBE), Frederic Mercier (Renault) and Olivier Roustant (ENSM-SE)

April 8, 2015



Outline

Outline

Problem description

Let us consider an expensive-to-evaluate black box simulator:

$$f : \mathcal{X} \subset \mathbb{R}^D \to G$$

The evaluation budget is very small, with $D \gg 1$.

Here $\mathcal{X} = [-1, 1]^D$, corresponding to box constraints.

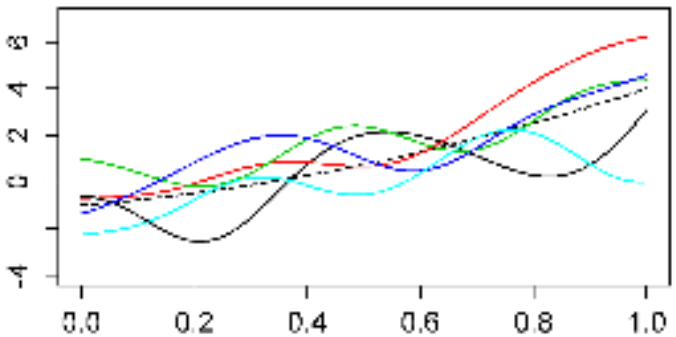
We want to optimize f when:

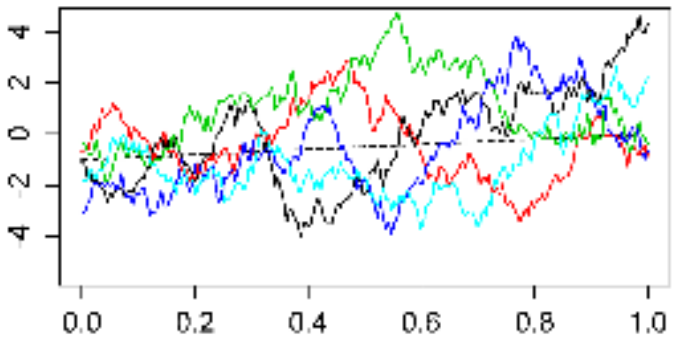
- $G = \mathbb{R}$, mono-objective case,
- $G = \mathbb{R}^m$, multi-objective (or constrained) case

Gaussian Process Regression - Kriging

Prior: f is considered as a realization of a Gaussian Process (GP) $\mathbf{Z}$.
A GP is fully determined by its mean and covariance kernel k functions.

- $\mathbf{Z} \sim \mathcal{GRF}(0, k(\cdot,\cdot))$ (*Simple Kriging*).
- $\mathbf{Z} \sim \mathcal{GRF}(g(\cdot,\beta), k(\cdot,\cdot))$ with $g(x,\beta) = \sum_{i-1}^{P} h_i(x)\beta_i$, the h_i are known basis functions (*Universal Kriging*).





Gaussian Process Regression - Kriging

Prior: f is considered as a realization of a Gaussian Process (GP) $\mathbf{Z}$.
A GP is fully determined by its mean and covariance kernel k functions.

- $\mathbf{Z} \sim \mathcal{GRF}(0, k(\cdot,\cdot))$ (*Simple Kriging*).
- $\mathbf{Z} \sim \mathcal{GRF}(g(\cdot,\beta), k(\cdot,\cdot))$ with $g(x,\beta) = \sum_{i=1}^{p} h_i(x)\beta_i$, the h_i are known basis functions (*Universal Kriging*).

Posterior: conditioning on n observations $\mathcal{A}_n = (Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n))$ $= (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$, the posterior distribution of $\mathbf{Z}$ at any point $\mathbf{x}$ is still Gaussian:

- Kriging mean: $m_n(\mathbf{x})$
- Kriging variance: $s_n^2(\mathbf{x})$
- $\mathcal{L}(Z(\mathbf{x})|\mathcal{A}_n) = \mathcal{N}(m_n(\mathbf{x}), s_n^2(\mathbf{x}))$

Bayesian optimization - Expected Improvement

Bayesian optimization

Sequential design strategy based on a distribution over functions to define an acquisition function (e.g. probability of improvement, Expected Improvement (EI) [JSW98], upper confidence bounds)

Bayesian optimization - Expected Improvement

Bayesian optimization

Sequential design strategy based on a distribution over functions to define an acquisition function (e.g. probability of improvement, Expected Improvement (EI) [JSW98], upper confidence bounds)

- Improvement defined as: $I : \begin{array}{ccc} \mathbf{D} & \rightarrow & \mathbb{R}^+ \\ u & \mapsto & max\left\{f_{min} - u, 0\right\} \end{array}$

  with f_{min} the current minimum.

Bayesian optimization - Expected Improvement

Bayesian optimization

Sequential design strategy based on a distribution over functions to define an acquisition function (e.g. probability of improvement, Expected Improvement (EI) [JSW98], upper confidence bounds)

- Improvement defined as: $I : \begin{array}{ccc} \mathbf{D} & \rightarrow & \mathbb{R}^+ \\ u & \mapsto & max\left\{f_{min} - u, 0\right\} \end{array}$

  with f_{min} the current minimum.

Expected Improvement

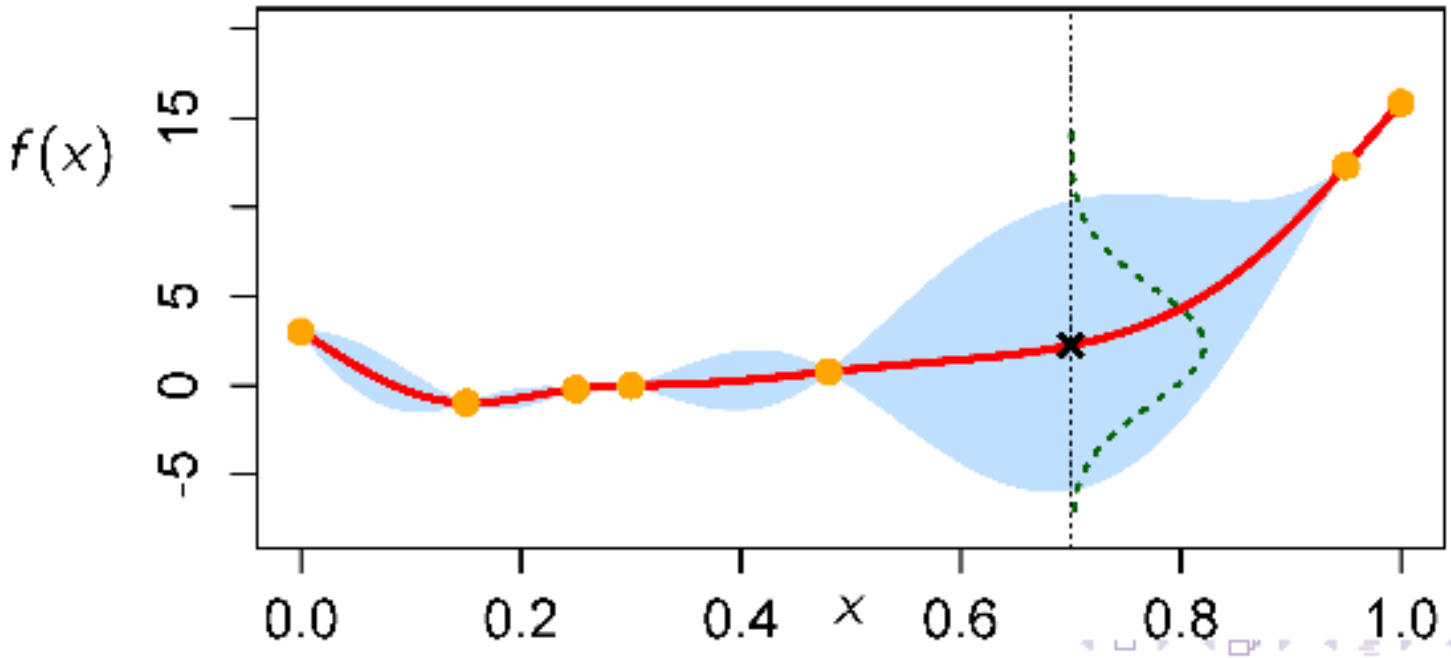
$$E[I(\mathbf{x})|\mathcal{A}_n] = (f_{min} - m_n(\mathbf{x}))\,\Phi\left(\frac{f_{min} - m_n(\mathbf{x})}{s_n(\mathbf{x})}\right) + s_n(\mathbf{x})\phi\left(\frac{f_{min} - m_n(\mathbf{x})}{s_n(\mathbf{x})}\right)$$

Bayesian optimization - Expected Improvement

Bayesian optimization

Sequential design strategy based on a distribution over functions to define an acquisition function (e.g. probability of improvement, Expected Improvement (EI) [JSW98], upper confidence bounds)

Example:

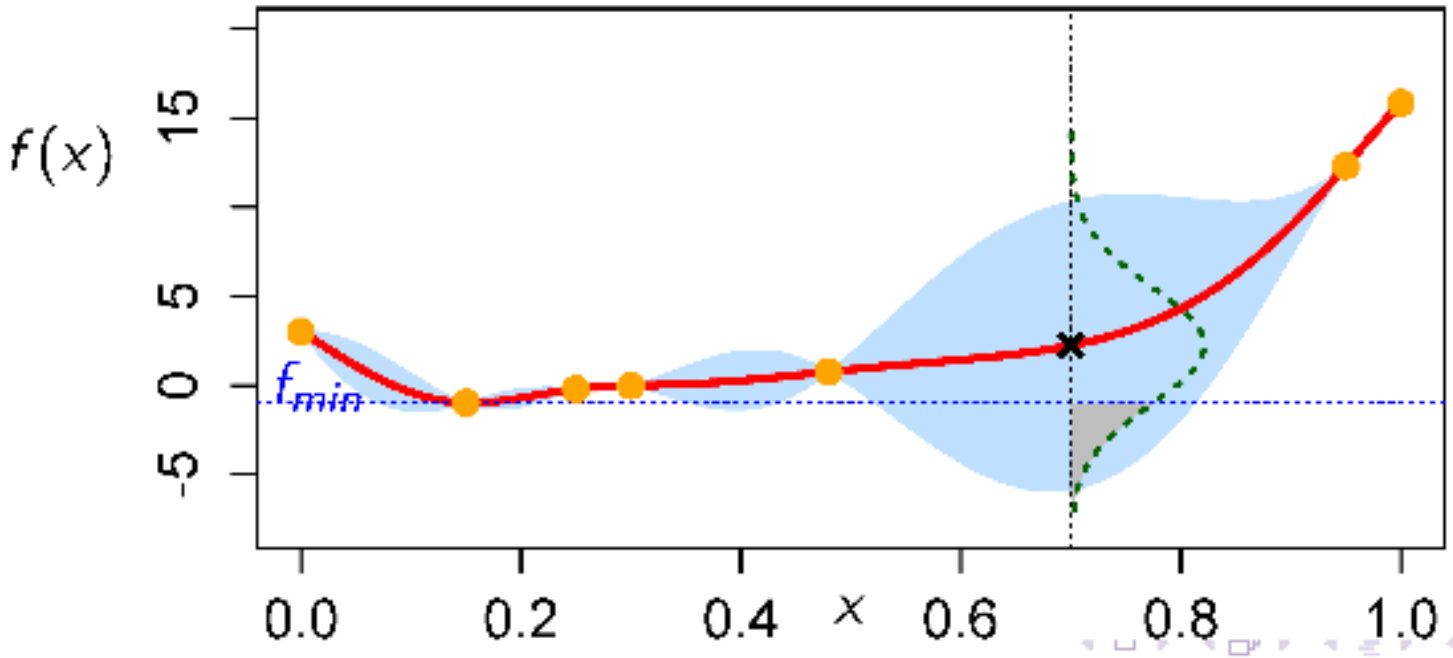


Bayesian optimization - Expected Improvement

Bayesian optimization

Sequential design strategy based on a distribution over functions to define an acquisition function (e.g. probability of improvement, Expected Improvement (EI) [JSW98], upper confidence bounds)

Example:

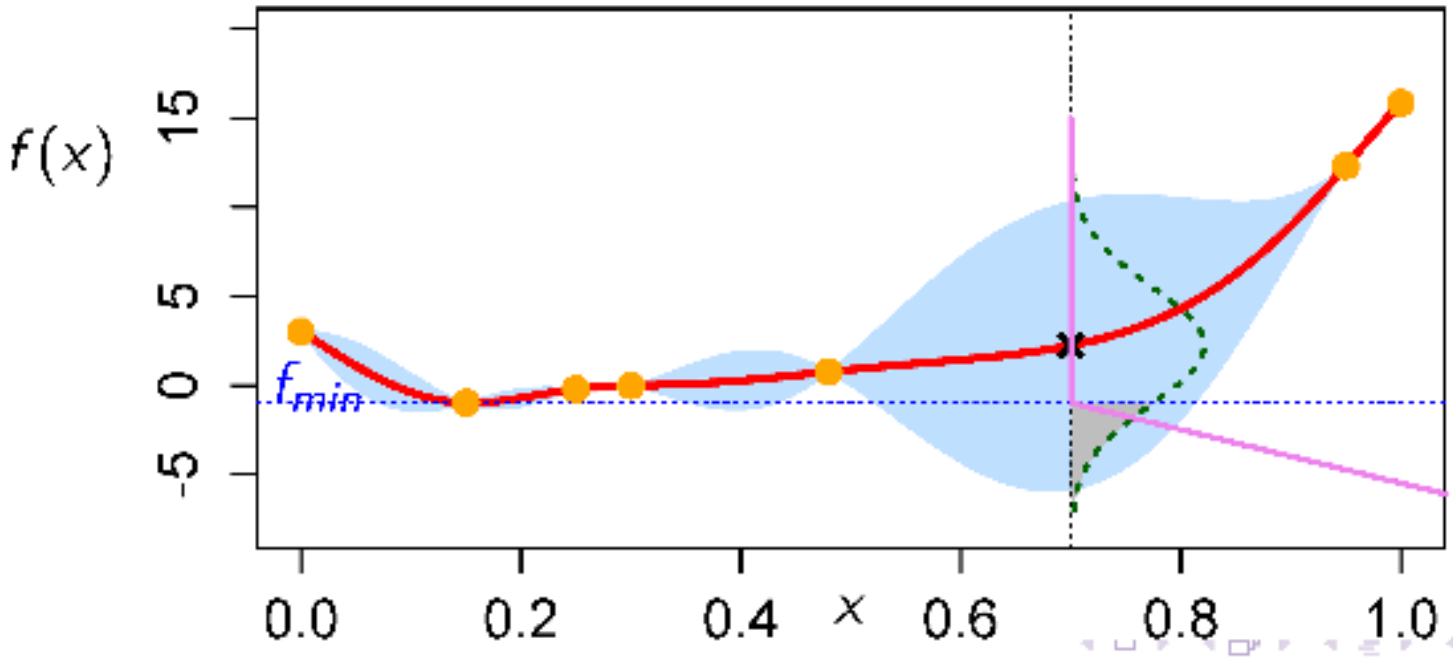


Bayesian optimization - Expected Improvement

Bayesian optimization

Sequential design strategy based on a distribution over functions to define an acquisition function (e.g. probability of improvement, Expected Improvement (EI) [JSW98], upper confidence bounds)

Example:

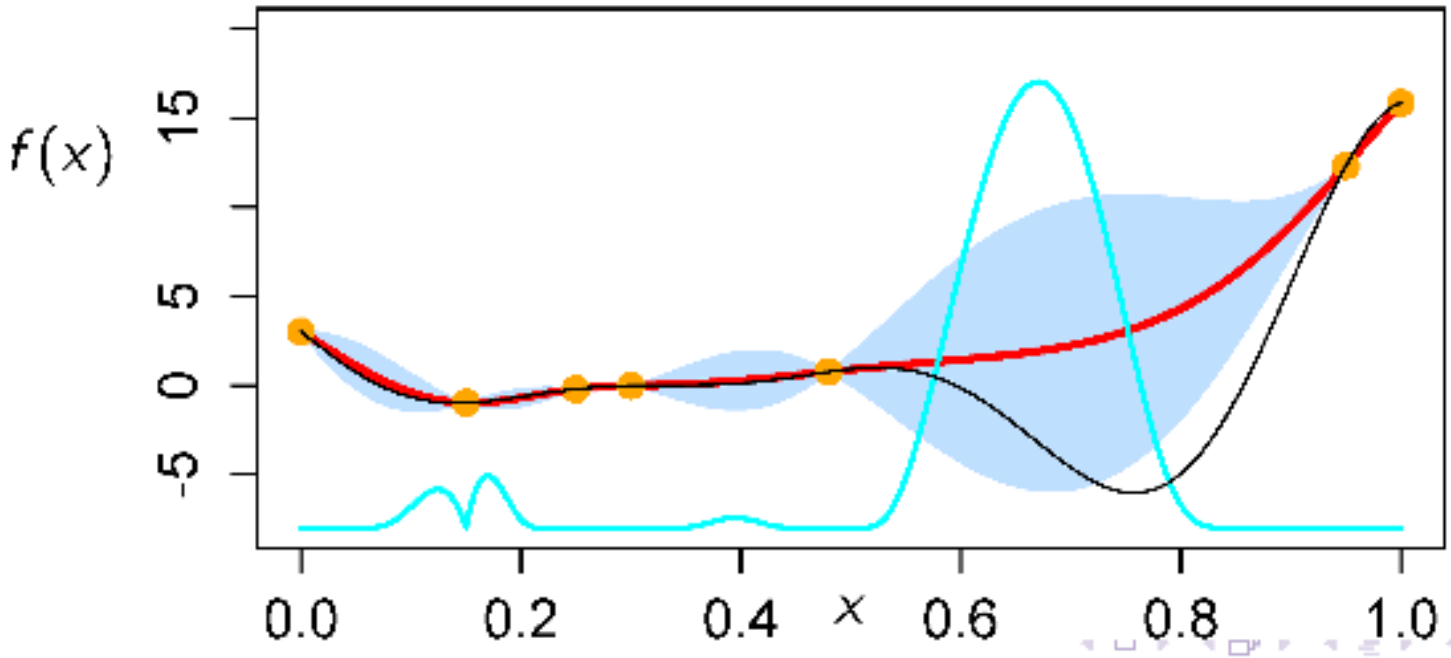


Bayesian optimization - Expected Improvement

Bayesian optimization

Sequential design strategy based on a distribution over functions to define an acquisition function (e.g. probability of improvement, Expected Improvement (EI) [JSW98], upper confidence bounds)

Example:

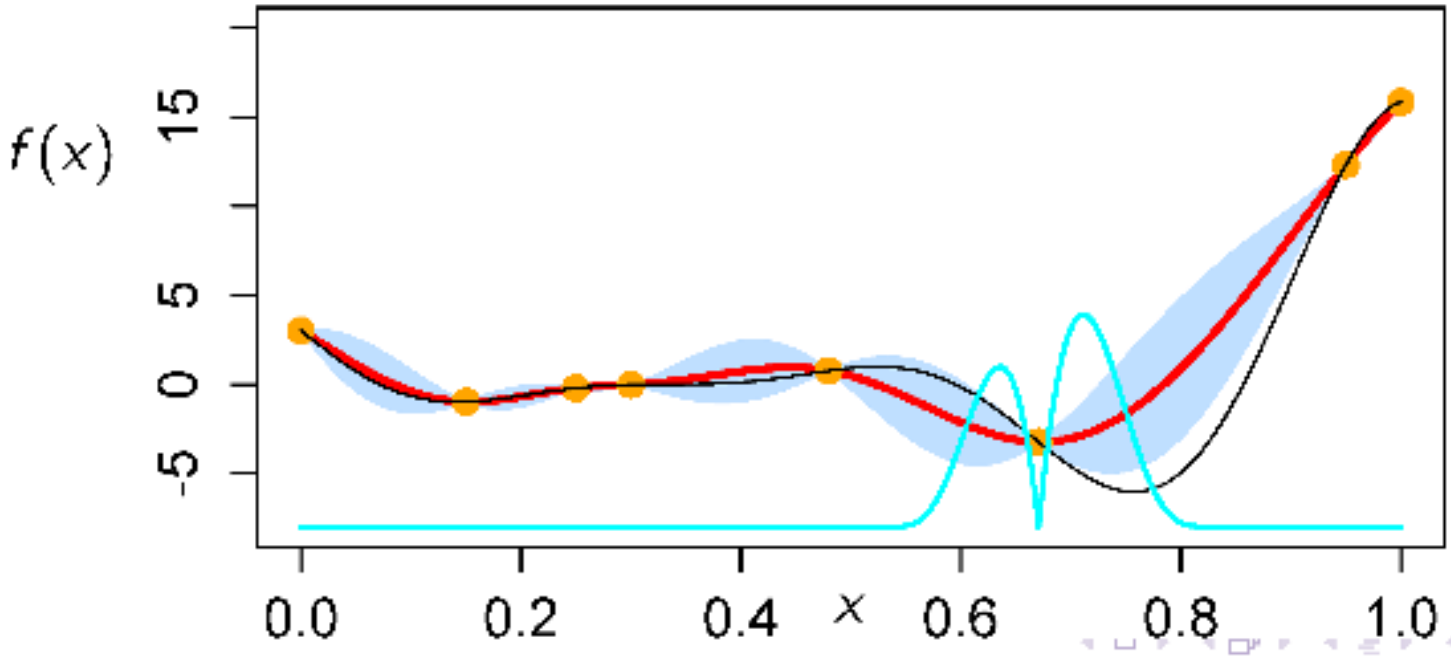


Bayesian optimization - Expected Improvement

Bayesian optimization

Sequential design strategy based on a distribution over functions to define an acquisition function (e.g. probability of improvement, Expected Improvement (EI) [JSW98], upper confidence bounds)

Example:

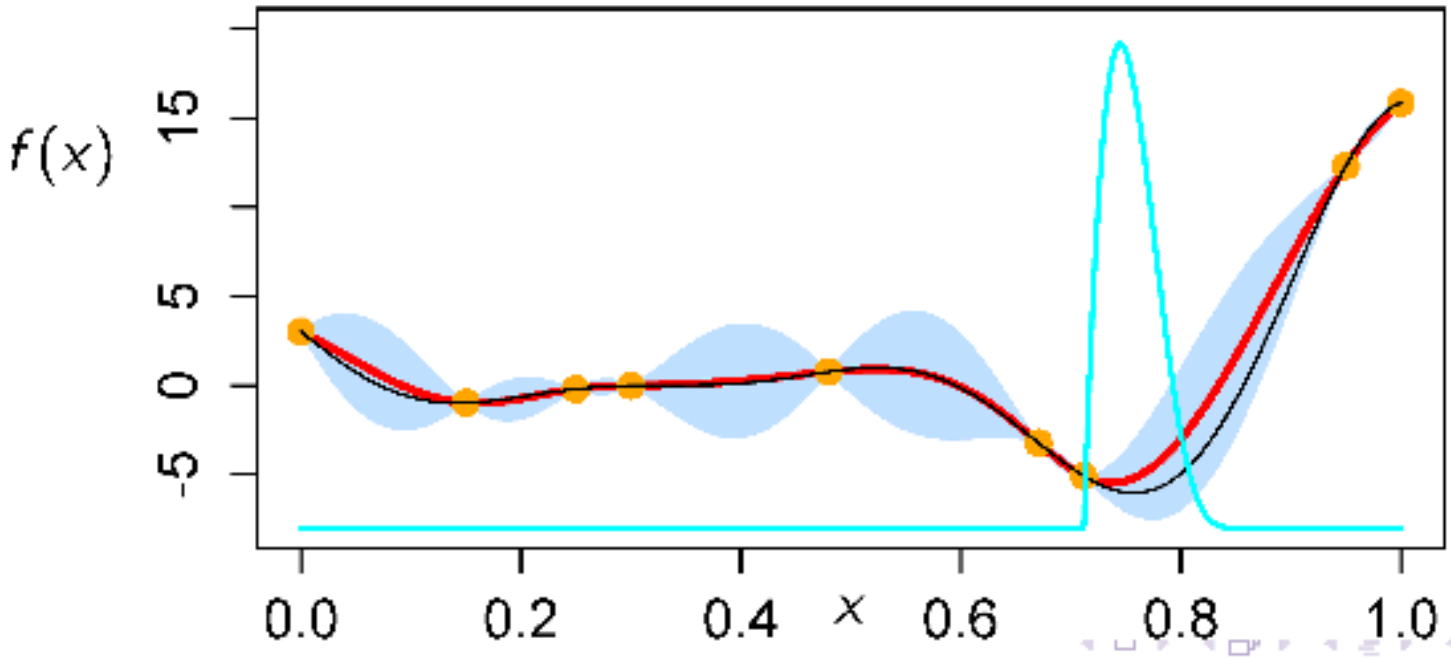


Bayesian optimization - Expected Improvement

Bayesian optimization

Sequential design strategy based on a distribution over functions to define an acquisition function (e.g. probability of improvement, Expected Improvement (EI) [JSW98], upper confidence bounds)

Example:



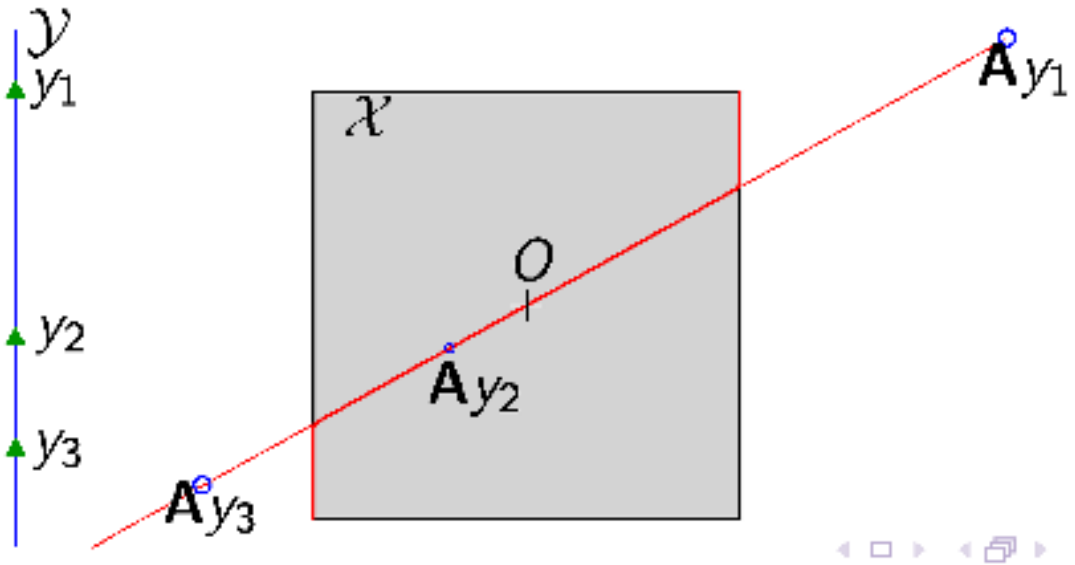
Random EMbedding Bayesian Optimization

Z. Wang, M. Zoghi, F. Hutter, D. Matheson, N. de Freitas
Bayesian Optimization in a High Dimensions via Random Embeddings
In: IJCAI 2013

Hypothesis: only a small number d_e of the D variables of f are influential.

Principle: using a random matrix $\mathbf{A} \in \mathbb{R}^{D \times d}$ with independent $\mathcal{N}(0,1)$ entries to map $\mathcal{Y} \subset \mathbb{R}^d$ onto $\mathcal{X}$, $d_e \leq d \ll D$.

A convex projection on $\mathcal{X}$, $p_{\mathcal{X}}$, is applied to points mapped outside of $\mathcal{X}$.



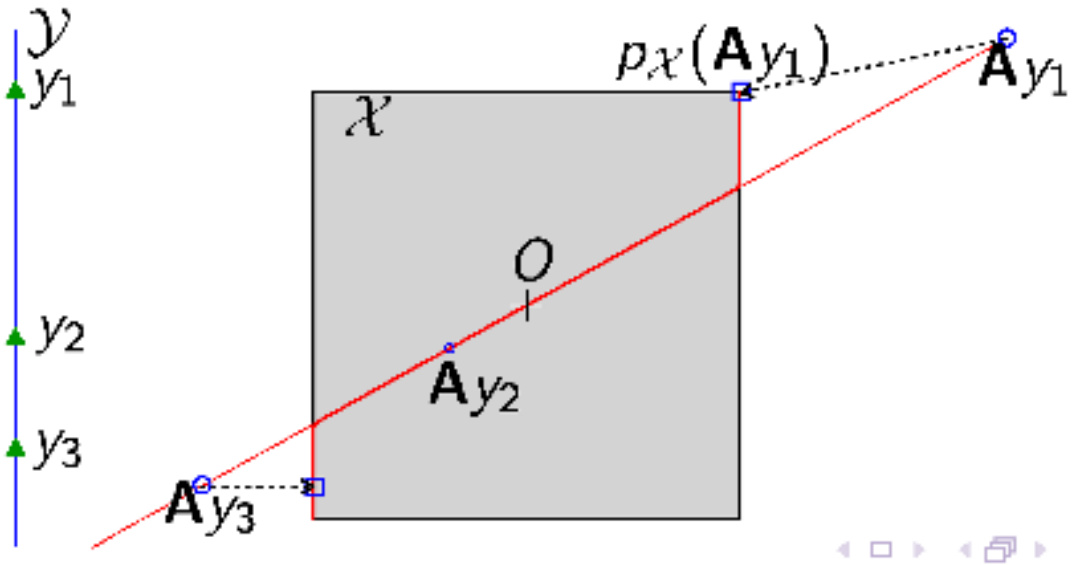
Random EMbedding Bayesian Optimization

Z. Wang, M. Zoghi, F. Hutter, D. Matheson, N. de Freitas
Bayesian Optimization in a High Dimensions via Random Embeddings
In: IJCAI 2013

Hypothesis: only a small number d_e of the D variables of f are influential.

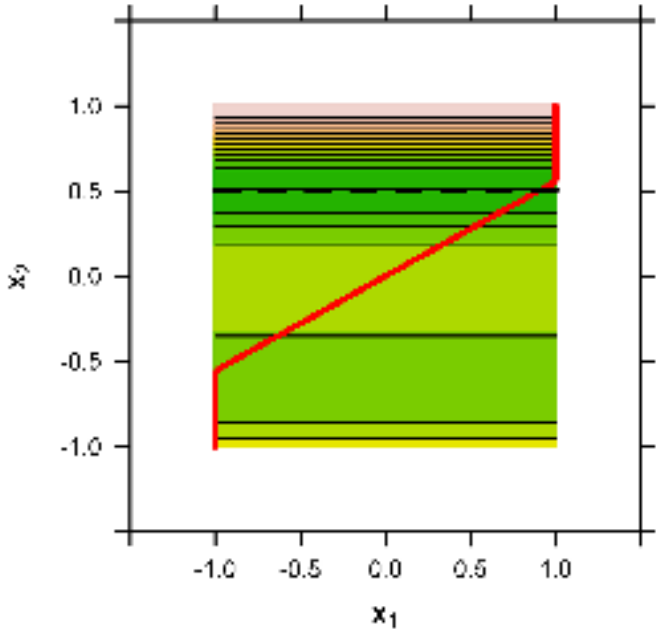
Principle: using a random matrix $\mathbf{A} \in \mathbb{R}^{D \times d}$ with independent $\mathcal{N}(0,1)$ entries to map $\mathcal{Y} \subset \mathbb{R}^d$ onto $\mathcal{X}$, $d_e \leq d \ll D$.

A convex projection on $\mathcal{X}$, $p_{\mathcal{X}}$, is applied to points mapped outside of $\mathcal{X}$.



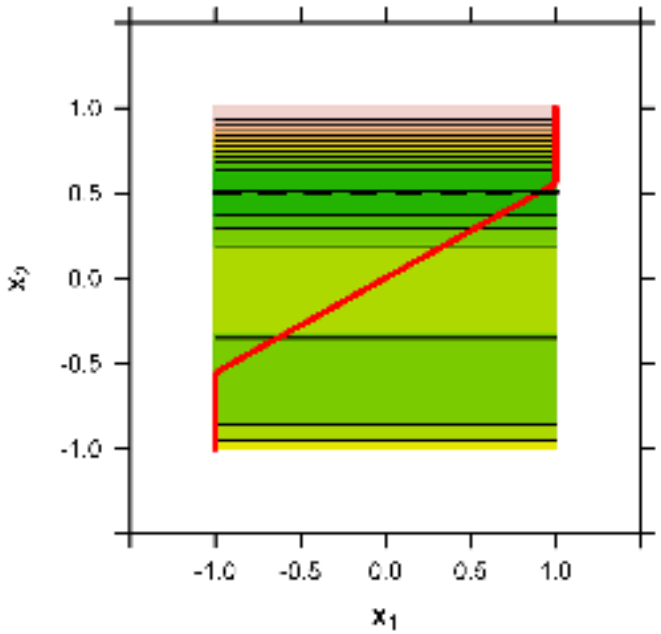
REMBO

With probability 1, a solution $\mathbf{y}^*$ corresponding to the optimum value can be found on the embedding.



REMBO

With probability 1, a solution $\mathbf{y}^*$ corresponding to the optimum value can be found on the embedding.



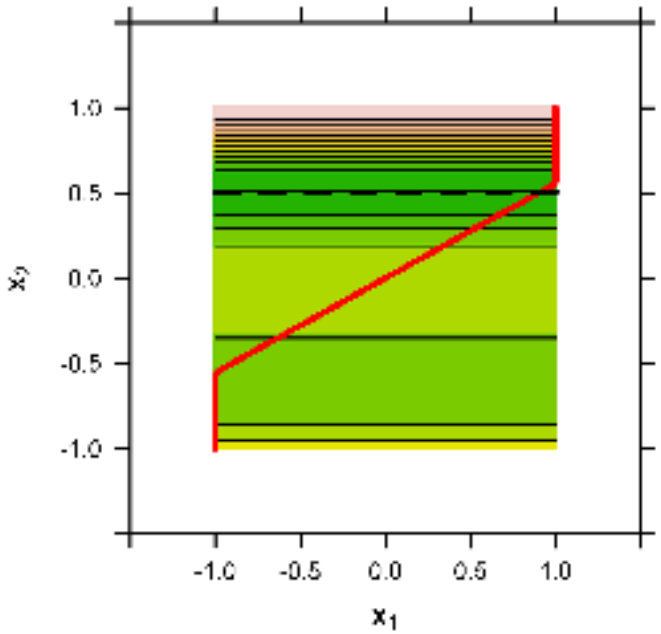
Bayesian optimization is performed on g :

$$\begin{array}{ccc} \mathcal{Y} \subset \mathbb{R}^d & \to & \mathbb{R} \\ \mathbf{y} & \mapsto & f(p_{\mathcal{X}}(\mathbf{A}\mathbf{y})) \end{array}$$

$\to$ Much smaller search space

REMBO

With probability 1, a solution $\mathbf{y}^*$ corresponding to the optimum value can be found on the embedding.



Bayesian optimization is performed on g :

$$\begin{array}{ccc} \mathcal{Y} \subset \mathbb{R}^d & \to & \mathbb{R} \\ \mathbf{y} & \mapsto & f(p_{\mathcal{X}}(\mathbf{A}\mathbf{y})) \end{array}$$

$\to$ Much smaller search space

How to best build the corresponding GP? In particular, with *which* kernel?

Choice of the covariance kernel for sequential optimization

Two covariance kernels are proposed in [WZM13]:

- kernel defined between points of $\mathcal{Y}$, for instance
  $k_{\mathcal{Y}}(\mathbf{y}, \mathbf{y'}) = exp(-\frac{\|\mathbf{y}-\mathbf{y}'\|_d^2}{2l^2})$
  - kriging in dimension d
  - suffers from the non-injectivity of the mapping

- kernel defined between projections in $\mathcal{X}$, for instance
  $k_{\mathcal{X}}(\mathbf{y}, \mathbf{y'}) = exp(-\frac{\|p_{\mathcal{X}}(\mathbf{Ay})-p_{\mathcal{X}}(\mathbf{Ay}')\|_D^2}{2l^2})$
  - kriging in dimension D
  - no problem with non-injectivity

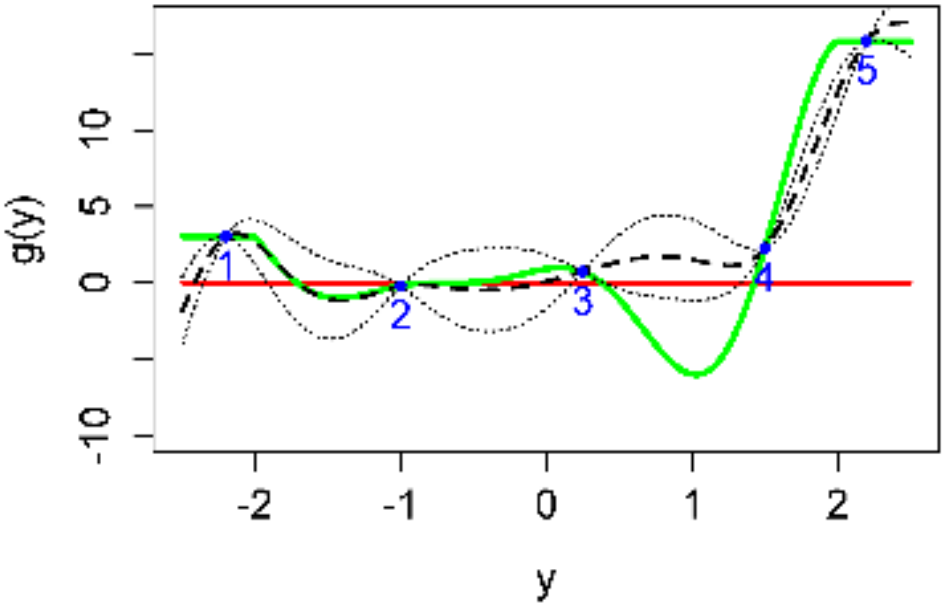
Illustration with $k_{\mathcal{Y}}$

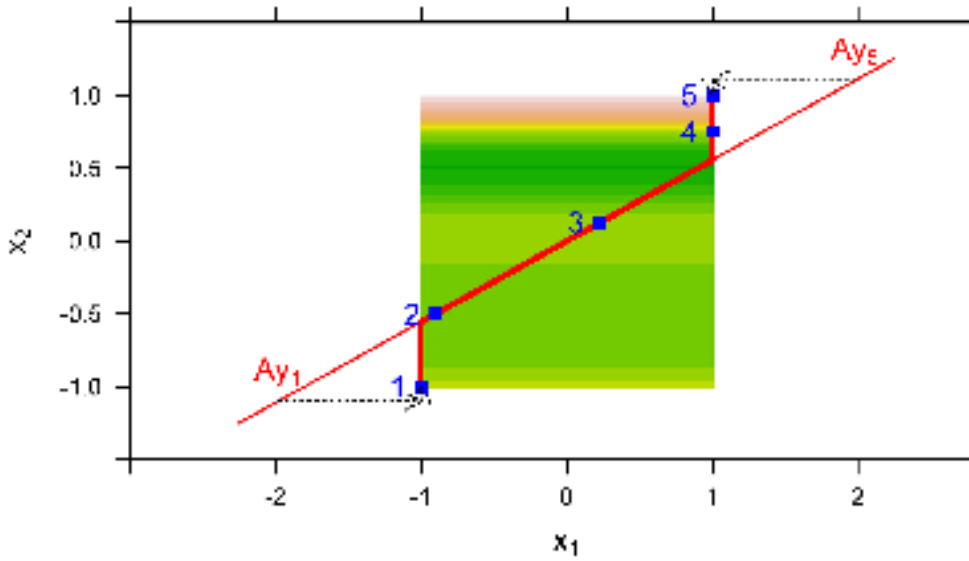
$$k_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') = exp(-\frac{\|\mathbf{y}-\mathbf{y}'\|_d^2}{2l^2})$$

$$\mathbf{A} = [0.9, 0.5]^T$$

$$\mathcal{Y} = [-2.5, 2.5]$$

$$\mathcal{X} = [-1, 1]^2$$





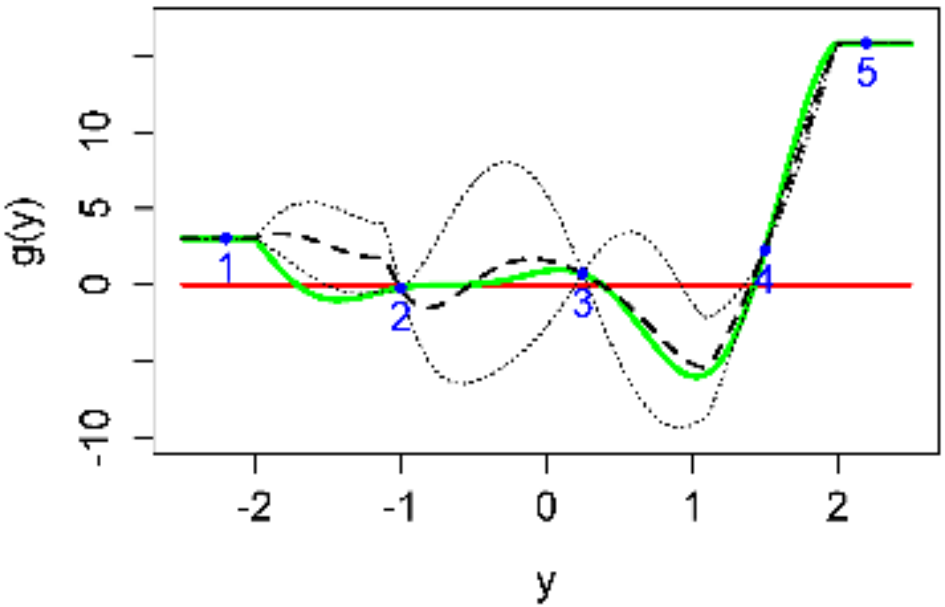
This kernel ignores that points in $\mathcal{Y}$ have the same projection onto $\mathcal{X}$.

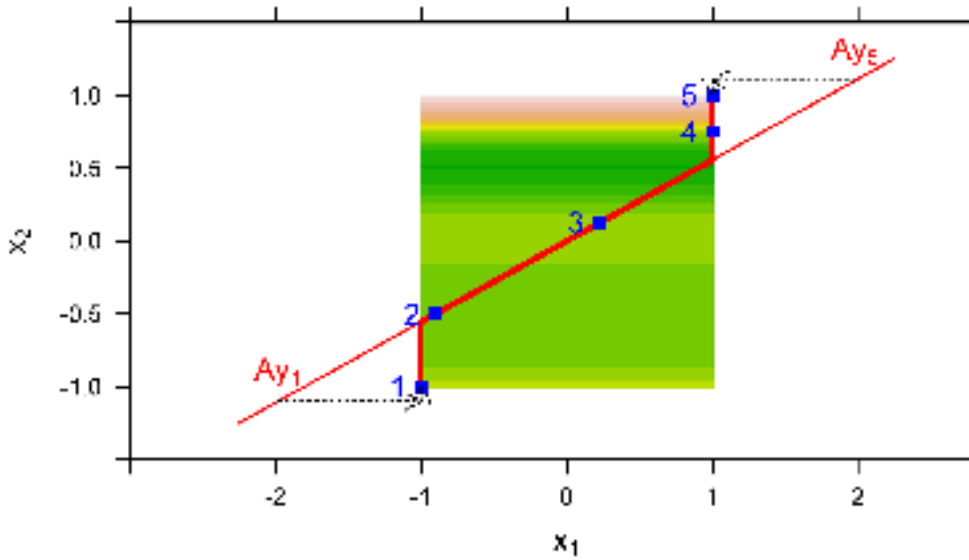
The bigger $\mathcal{Y}$, the more the exploration of its sides will be important.

Illustration with $k_{\mathcal{X}}$

$$k_{\mathcal{X}}(\mathbf{y}, \mathbf{y}') = exp(-\frac{\|p_{\mathcal{X}}(\mathbf{Ay}) - p_{\mathcal{X}}(\mathbf{Ay}')\|_D^2}{2l^2}) \qquad \mathbf{A} = [0.9, 0.5]^T$$

$$\mathcal{Y} = [-2.5, 2.5] \qquad \mathcal{X} = [-1, 1]^2$$





Here, no problem with already known points in $\mathcal{X}$.

But $k_{\mathcal{X}}$ is D-dimensional, thus impacted by the curse of dimensionality, except inside of $\mathcal{X}$.

Selecting the size of $\mathcal{Y}$

It can be shown that with probability $(1-\epsilon)$, $\|\mathbf{y}^*\|_2 \leq \frac{d_e}{\epsilon}$ [WZM13].
$\rightarrow \mathcal{Y}$ should be taken as large as possible to be sure to find the optimum.

But taking $\mathcal{Y}$ too large:

- causes problems with $k_{\mathcal{Y}}$ or $k_{\mathcal{X}}$
- renders the optimization of the Expected Improvement intractable

Selecting the size of $\mathcal{Y}$

It can be shown that with probability $(1-\epsilon)$, $\|\mathbf{y}^*\|_2 \leq \frac{d_e}{\epsilon}$ [WZM13].
$\rightarrow \mathcal{Y}$ should be taken as large as possible to be sure to find the optimum.

But taking $\mathcal{Y}$ too large:

- causes problems with $k_{\mathcal{Y}}$ or $k_{\mathcal{X}}$
- renders the optimization of the Expected Improvement intractable

As a compromise, $\mathcal{Y}$ is set to $[-\sqrt{d}, \sqrt{d}]^d$ and either

- d is defined such that $d > d_e$
- the evaluation budget is split between several random embeddings

Selecting the size of $\mathcal{Y}$

It can be shown that with probability $(1-\epsilon)$, $\|\mathbf{y}^*\|_2 \leq \frac{d_e}{c}$ [WZM13].
$\rightarrow$ $\mathcal{Y}$ should be taken as large as possible to be sure to find the optimum.

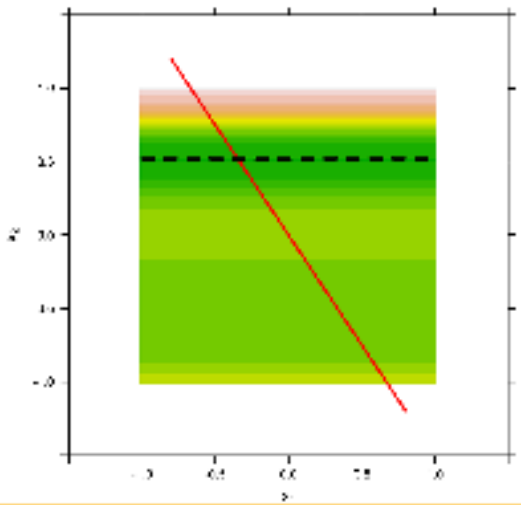
But taking $\mathcal{Y}$ too large:

- causes problems with $k_{\mathcal{Y}}$ or $k_{\mathcal{X}}$
- renders the optimization of the Expected Improvement intractable

As a compromise, $\mathcal{Y}$ is set to $[-\sqrt{d}, \sqrt{d}]^d$ and either

- d is defined such that $d > d_e$
- the evaluation budget is split between several random embeddings

Fixed size $\mathcal{Y}$:



Selecting the size of $\mathcal{Y}$

It can be shown that with probability $(1-\epsilon)$, $\|\mathbf{y}^*\|_2 \leq \frac{d_e}{c}$ [WZM13].
$\rightarrow$ $\mathcal{Y}$ should be taken as large as possible to be sure to find the optimum.

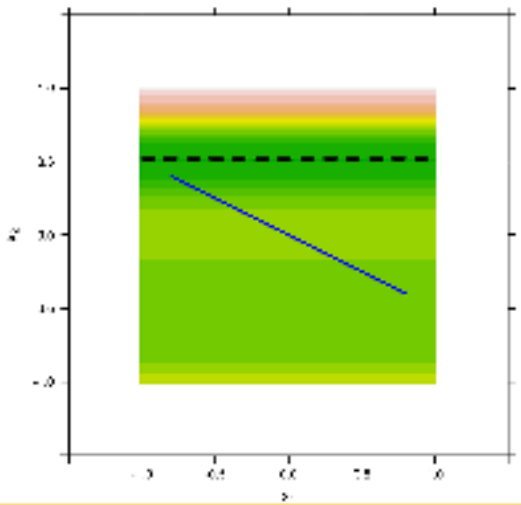
But taking $\mathcal{Y}$ too large:

- causes problems with $k_{\mathcal{Y}}$ or $k_{\mathcal{X}}$
- renders the optimization of the Expected Improvement intractable

As a compromise, $\mathcal{Y}$ is set to $[-\sqrt{d}, \sqrt{d}]^d$ and either

- d is defined such that $d > d_e$
- the evaluation budget is split between several random embeddings

Fixed size $\mathcal{Y}$:



Selecting the size of $\mathcal{Y}$

It can be shown that with probability $(1-\epsilon)$, $\|\mathbf{y}^*\|_2 \leq \frac{d_e}{c}$ [WZM13].
$\rightarrow$ $\mathcal{Y}$ should be taken as large as possible to be sure to find the optimum.

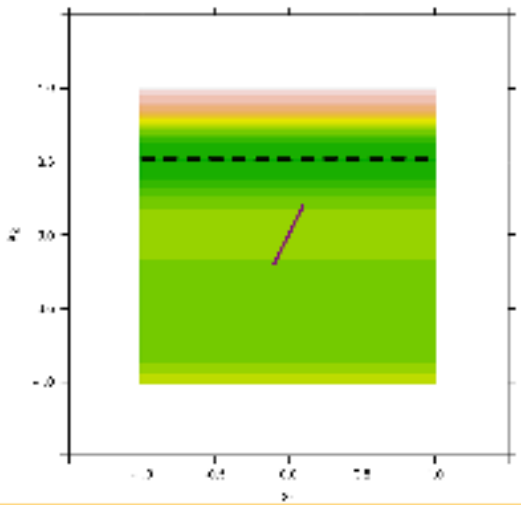
But taking $\mathcal{Y}$ too large:

- causes problems with $k_{\mathcal{Y}}$ or $k_{\mathcal{X}}$
- renders the optimization of the Expected Improvement intractable

As a compromise, $\mathcal{Y}$ is set to $[-\sqrt{d}, \sqrt{d}]^d$ and either

- d is defined such that $d > d_e$
- the evaluation budget is split between several random embeddings

Fixed size $\mathcal{Y}$:



Selecting the size of $\mathcal{Y}$

It can be shown that with probability $(1-\epsilon)$, $\|\mathbf{y}^*\|_2 \leq \frac{d_e}{c}$ [WZM13].
$\rightarrow$ $\mathcal{Y}$ should be taken as large as possible to be sure to find the optimum.

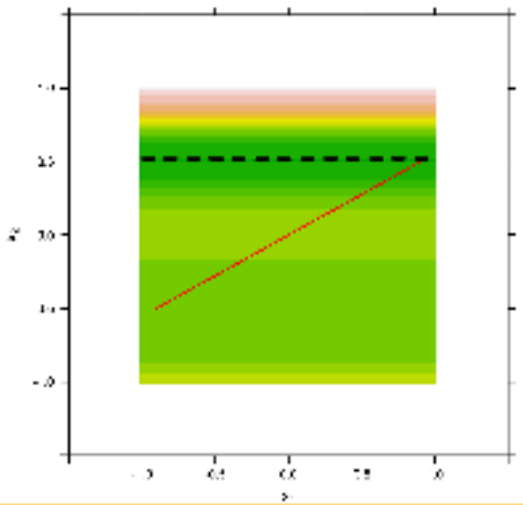
But taking $\mathcal{Y}$ too large:

- causes problems with $k_{\mathcal{Y}}$ or $k_{\mathcal{X}}$
- renders the optimization of the Expected Improvement intractable

As a compromise, $\mathcal{Y}$ is set to $[-\sqrt{d}, \sqrt{d}]^d$ and either

- d is defined such that $d > d_e$
- the evaluation budget is split between several random embeddings

Fixed size $\mathcal{Y}$:



Outline

Motivations - derivation

Objective: Creating a kernel with the benefits of both $k_{\mathcal{Y}}$ and $k_{\mathcal{X}}$:

- not affected by non-injectivity
- accounting for high-dimensional correlations
- working in a low-dimensional space
- scalable with high D and with moderate d

Motivations - derivation

Objective: Creating a kernel with the benefits of both $k_{\mathcal{Y}}$ and $k_{\mathcal{X}}$:

- not affected by non-injectivity
- accounting for high-dimensional correlations
- working in a low-dimensional space
- scalable with high D and with moderate d

Remark: $k_{\mathcal{X}}(\mathbf{y}, \mathbf{y}') = \exp(-\|\mathbf{u}(\mathbf{y}) - \mathbf{u}(\mathbf{y}')\|_D^2 / 2l^2)$ where $\mathbf{u} : \mathcal{Y} \to \mathcal{X}$, $\mathbf{u}(\mathbf{y}) = p_{\mathcal{X}}(\mathbf{A}\mathbf{y})$ is a *warping* of the input space $\mathcal{Y}$ [Ras06].

Motivations - derivation

Objective: Creating a kernel with the benefits of both $k_{\mathcal{Y}}$ and $k_{\mathcal{X}}$:

- not affected by non-injectivity
- accounting for high-dimensional correlations
- working in a low-dimensional space
- scalable with high D and with moderate d

Remark: $k_{\mathcal{X}}(\mathbf{y}, \mathbf{y}') = \exp(-\|\mathbf{u}(\mathbf{y}) - \mathbf{u}(\mathbf{y}')\|_D^2 / 2l^2)$ where $\mathbf{u} : \mathcal{Y} \to \mathcal{X}$, $\mathbf{u}(\mathbf{y}) = p_{\mathcal{X}}(\mathbf{A}\mathbf{y})$ is a *warping* of the input space $\mathcal{Y}$ [Ras06].

Our proposed kernel integrates a modified version of this warping with:

1. Back projection onto $\mathrm{Ran}(\mathbf{A})$ of points of the warped points, using the orthogonal projection $p_{\mathbf{A}} : \mathcal{X} \mapsto \mathbb{R}^D$, $p_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}$

Motivations - derivation

Objective: Creating a kernel with the benefits of both $k_{\mathcal{Y}}$ and $k_{\mathcal{X}}$:

- not affected by non-injectivity
- accounting for high-dimensional correlations
- working in a low-dimensional space
- scalable with high D and with moderate d

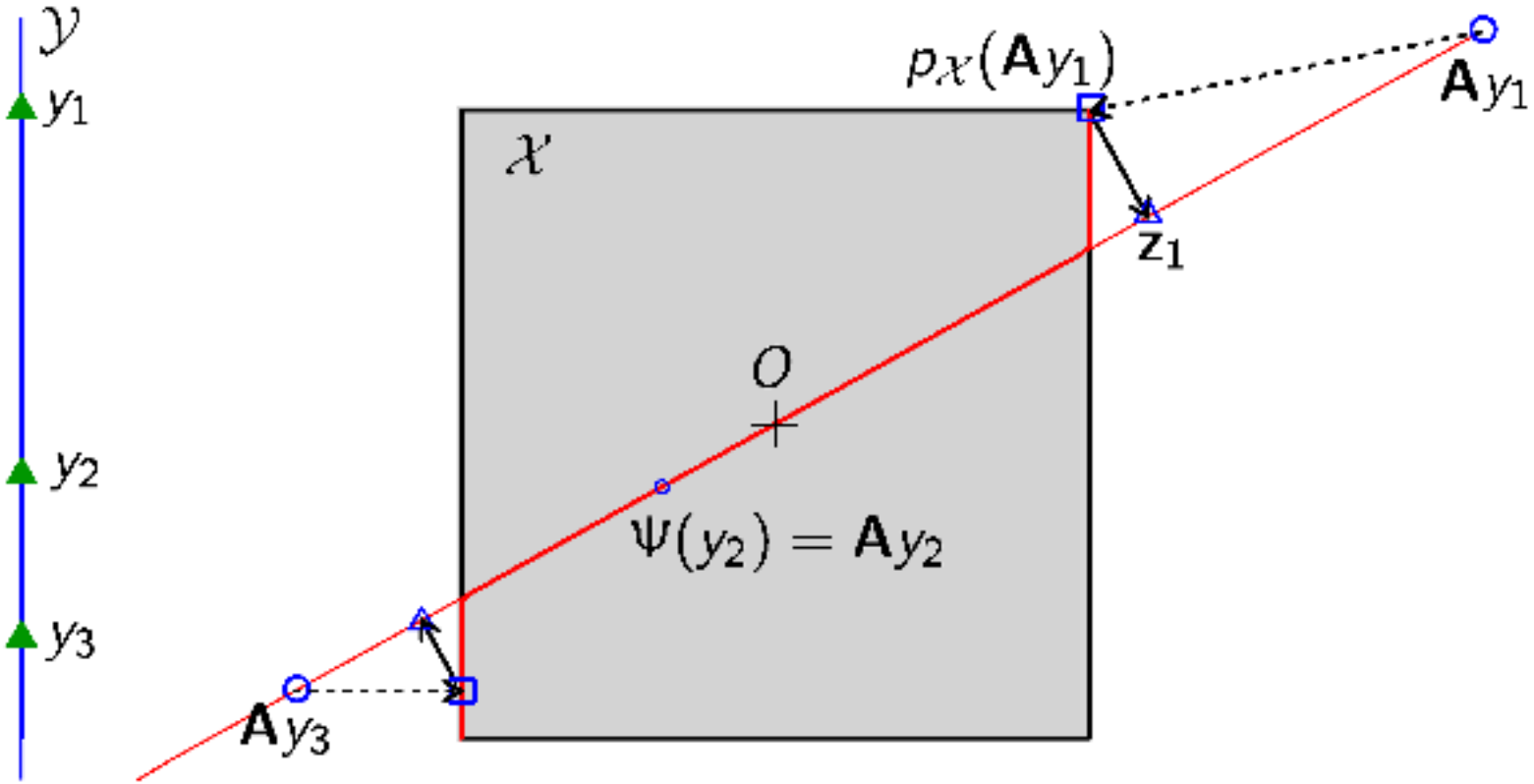
Remark: $k_{\mathcal{X}}(\mathbf{y}, \mathbf{y}') = \exp(-\|\mathbf{u}(\mathbf{y}) - \mathbf{u}(\mathbf{y}')\|_D^2 / 2l^2)$ where $\mathbf{u} : \mathcal{Y} \to \mathcal{X}$, $\mathbf{u}(\mathbf{y}) = p_{\mathcal{X}}(\mathbf{A}\mathbf{y})$ is a *warping* of the input space $\mathcal{Y}$ [Ras06].

Our proposed kernel integrates a modified version of this warping with:

1. Back projection onto $\mathrm{Ran}(\mathbf{A})$ of points of the warped points, using the orthogonal projection $p_{\mathbf{A}} : \mathcal{X} \mapsto \mathbb{R}^D$, $p_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x}$
2. Distortion of those back projection to avoid possible instationarity problems, from a pivot point

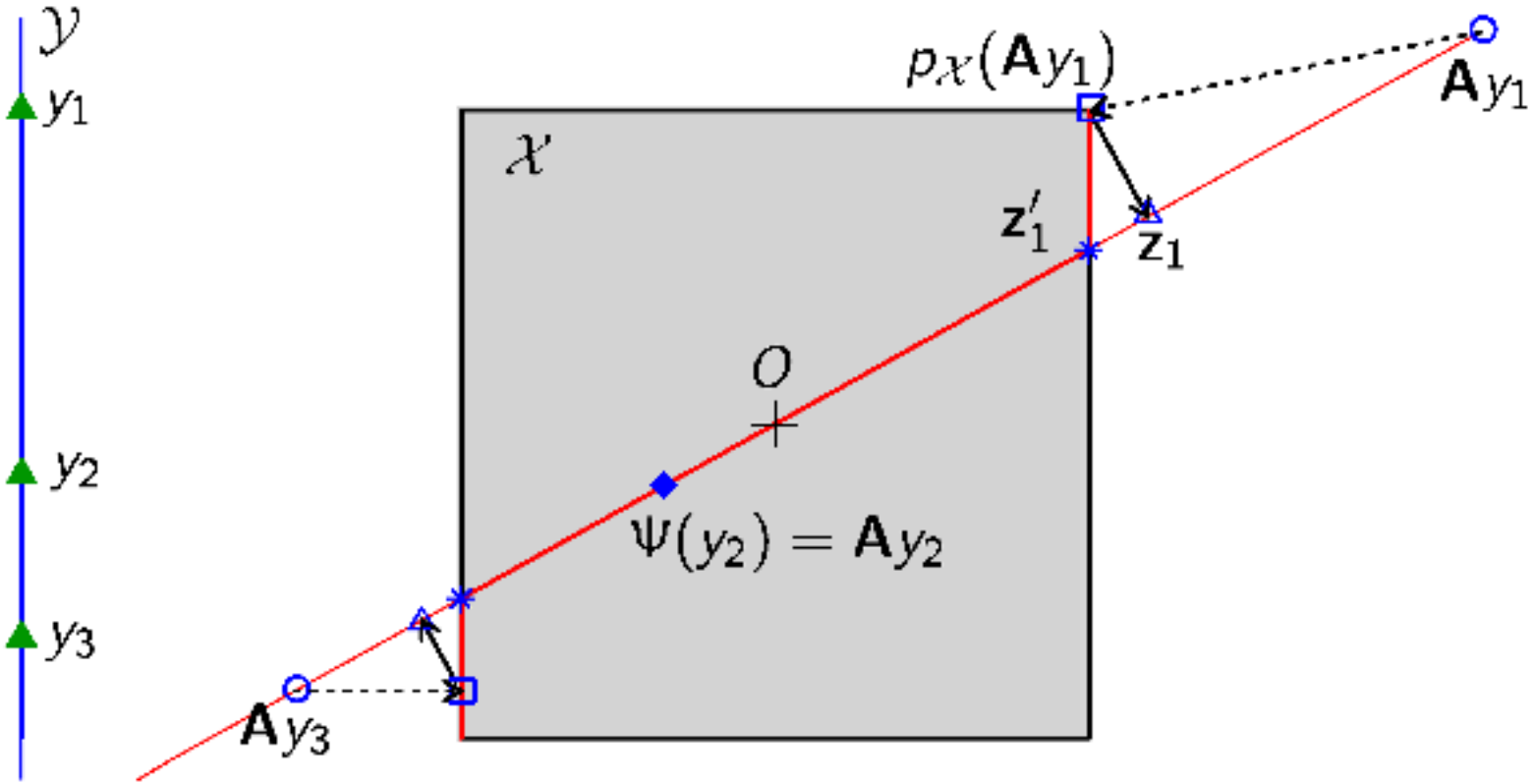
Proposition of warping

Retro-projection in a low dimensional space:
Orthogonal projection onto Ran($\mathbf{A}$) with $p_{\mathbf{A}}$



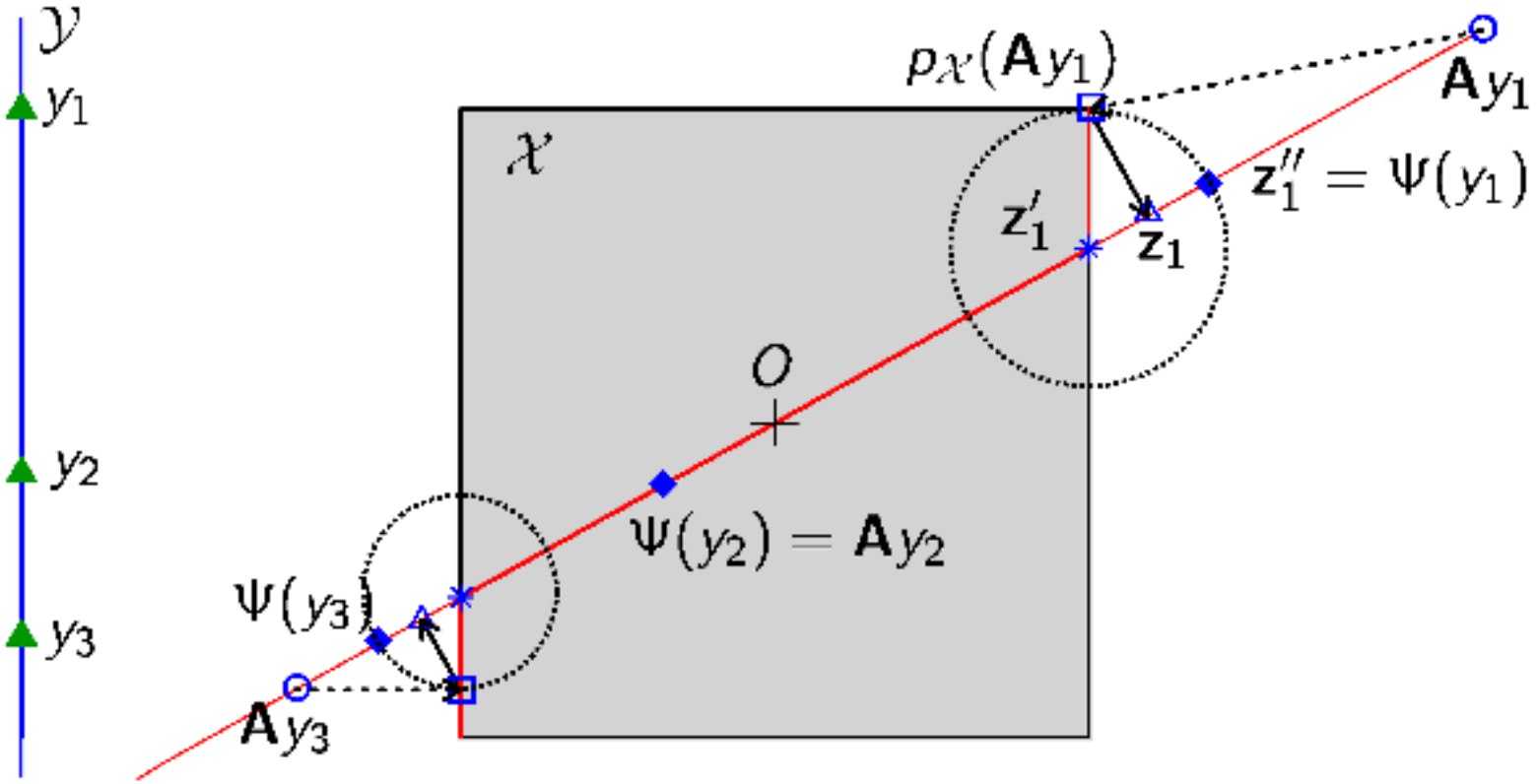
Proposition of warping

Problem: loss of high-dimensional distance information
(on convex projected parts)



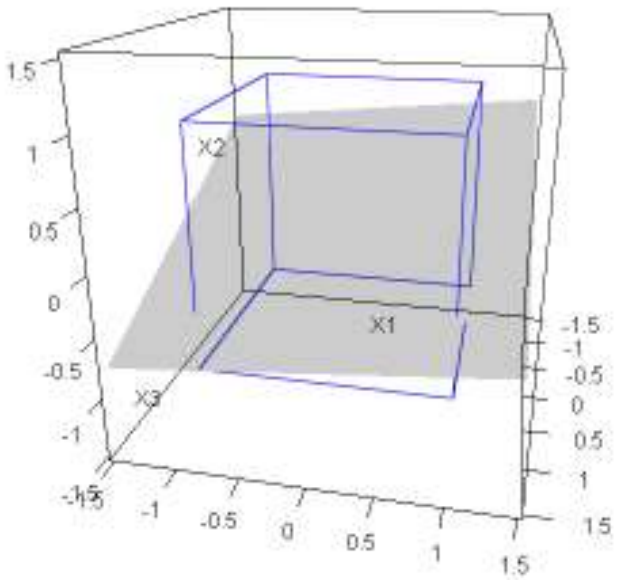
Proposition of warping

Distortion to account for high-dimensional distance **on** the embedded space



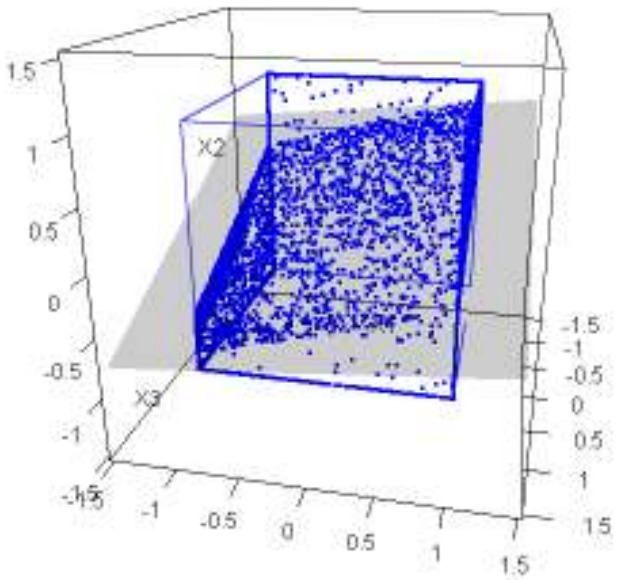
Motivation for the distortion

Example ($d = 2$, $D = 3$)



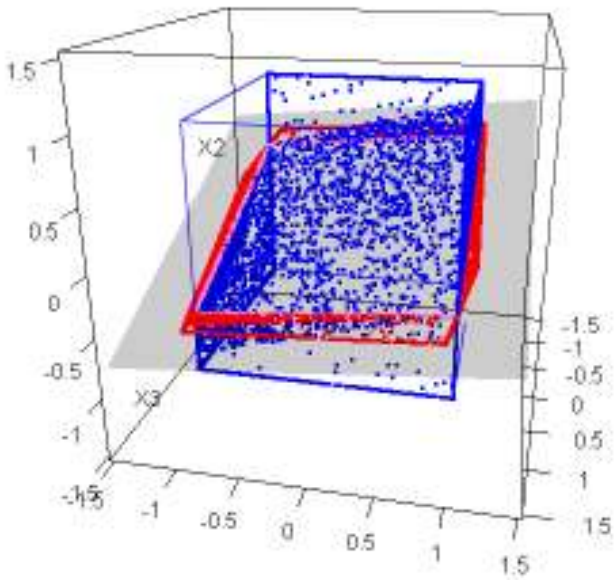
Motivation for the distortion

Example ($d = 2$, $D = 3$)



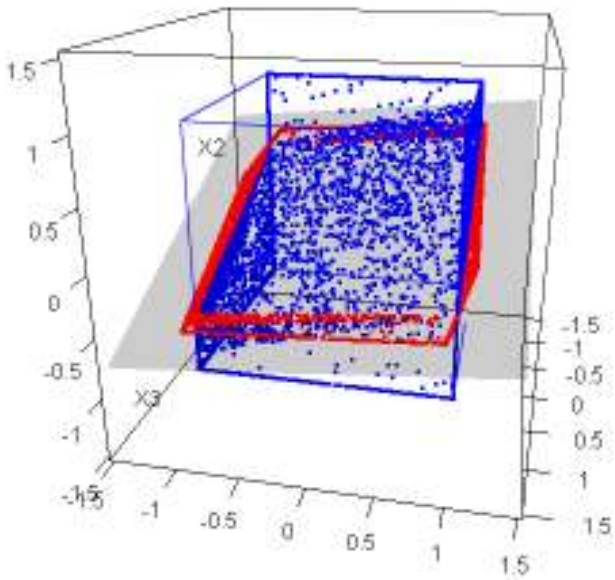
Motivation for the distortion

Example ($d = 2$, $D = 3$)



Motivation for the distortion

Example ($d = 2$, $D = 3$)



Motivation for the distortion

Example ($d = 2$, $D = 3$)

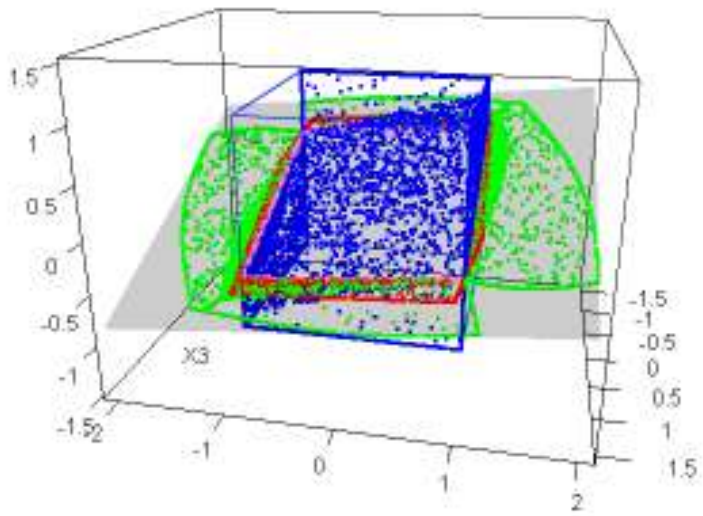
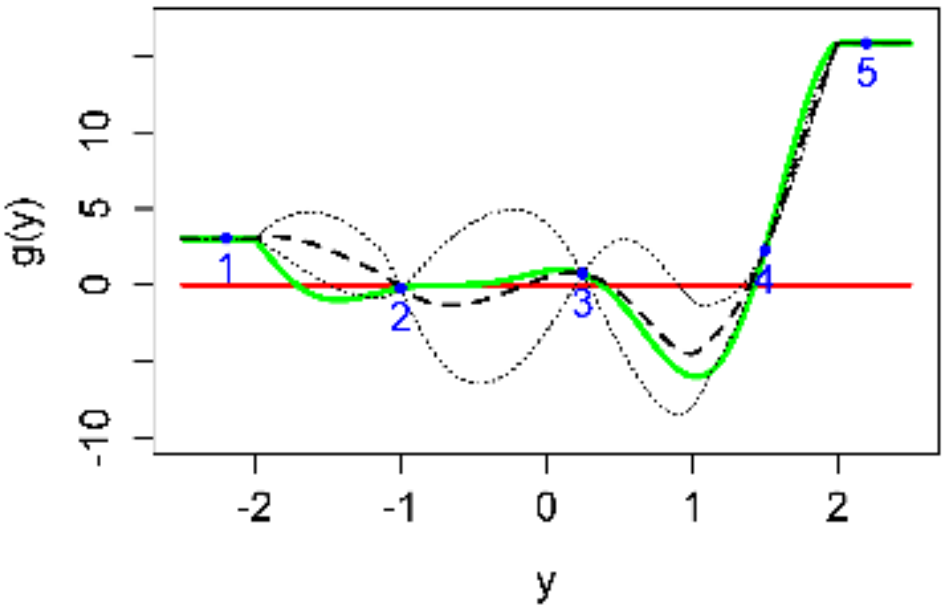


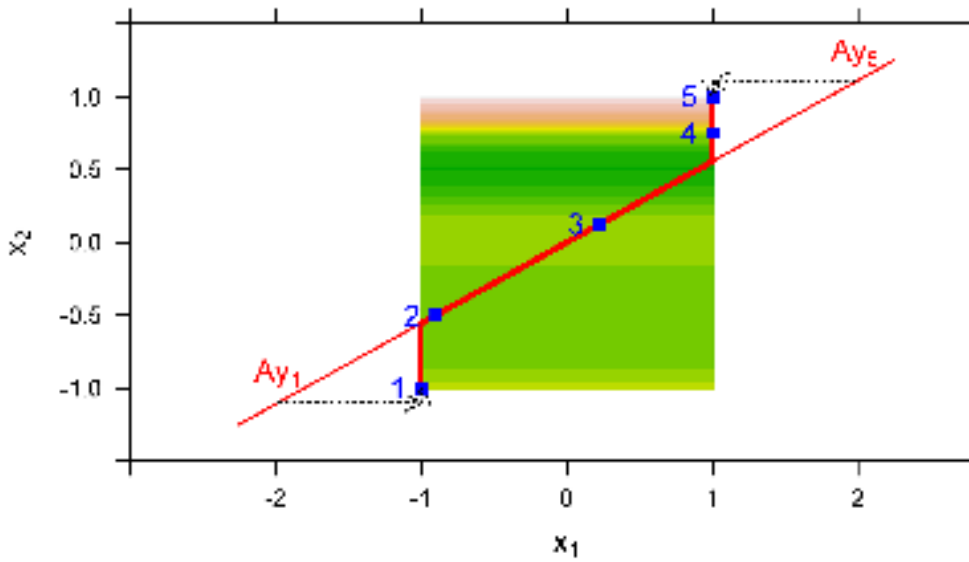
Illustration with k_ψ

$$\mathbf{A} = [0.9, 0.5]^T$$

$\mathcal{Y} = [-2.5, 2.5]$



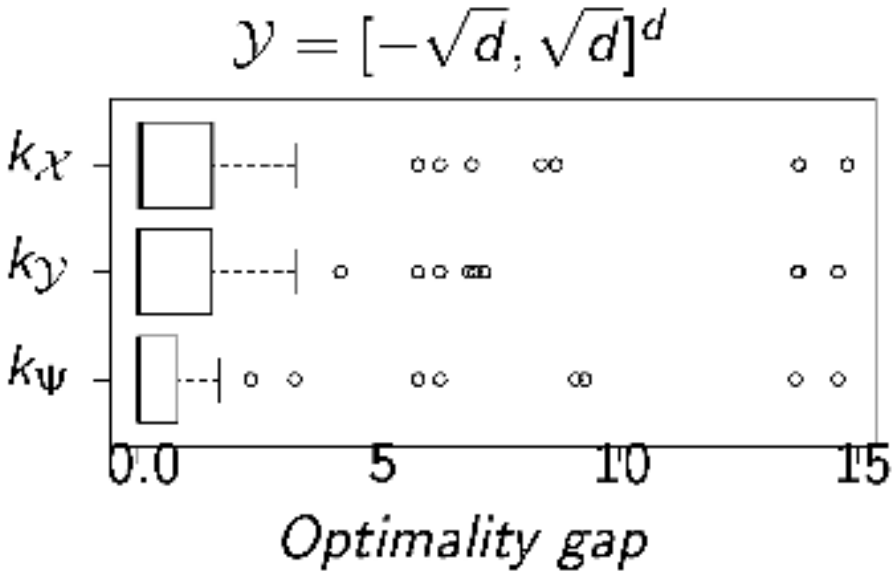
$\mathcal{X} = [-1, 1]^2$

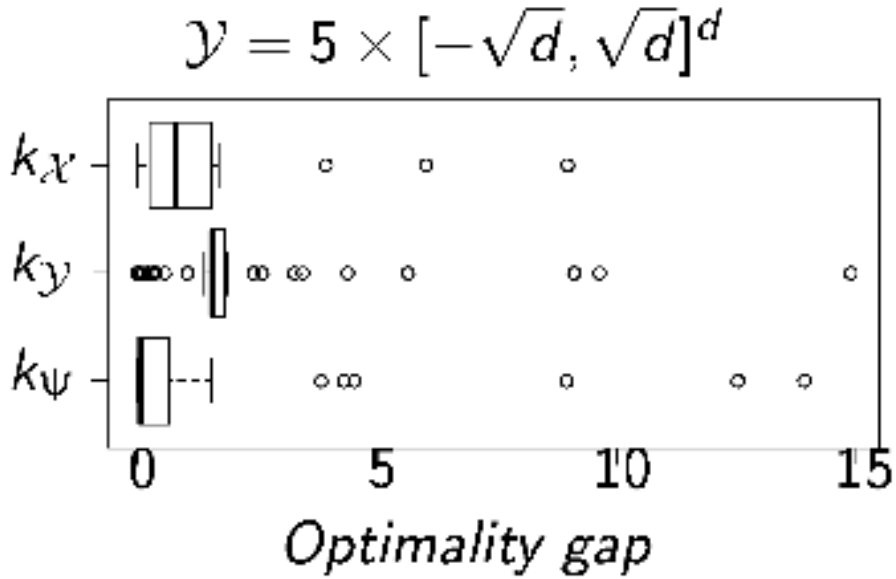


No problem with already known points **and** not suffering from the curse of dimensionality.

Results on the Branin test function

100 evaluations, $d = 2$, $D = 25$, 50 repetitions:

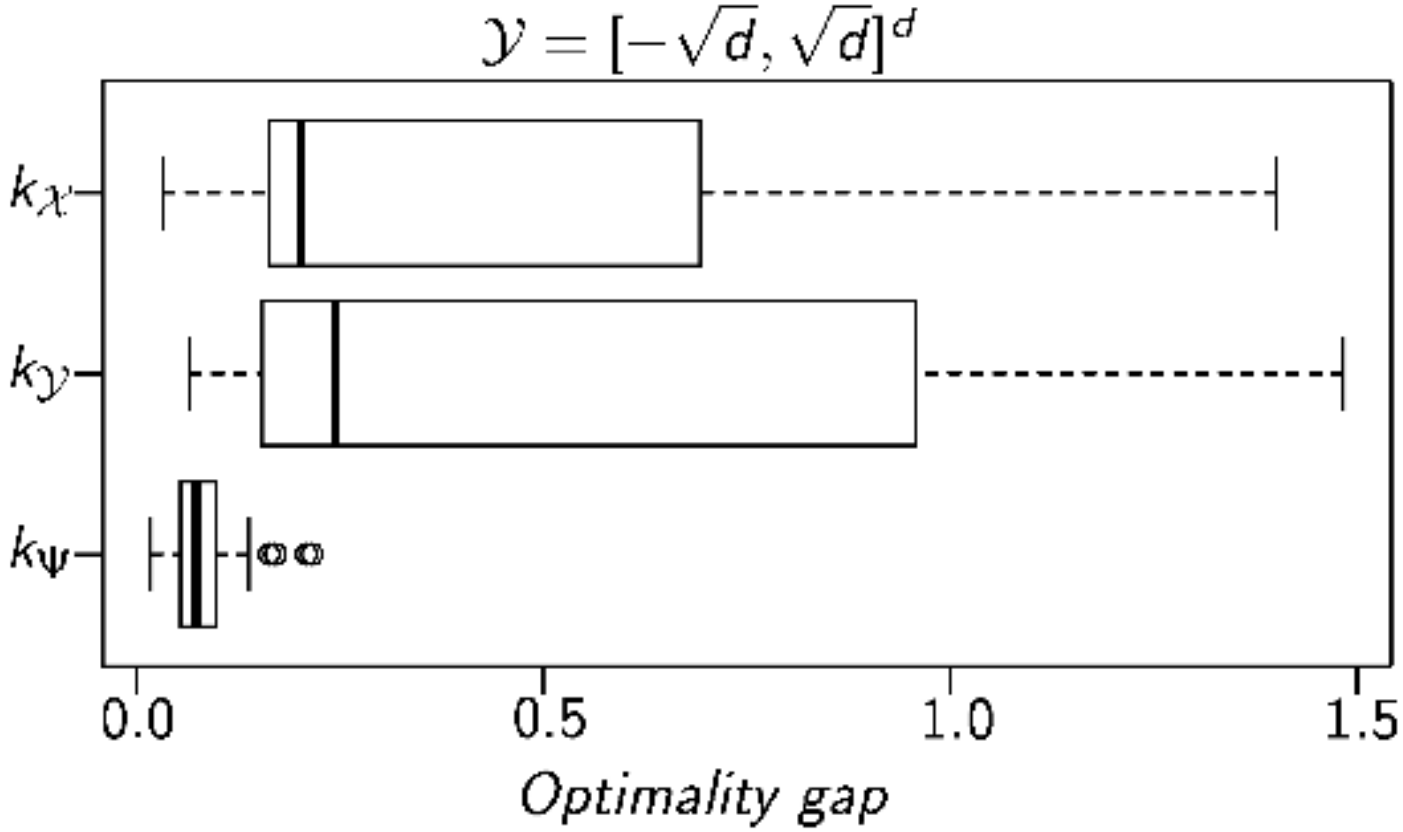




k_{ψ} has the best performance and is more robust when the size of $\mathcal{Y}$ increases.

Results on the Hartman6 test function

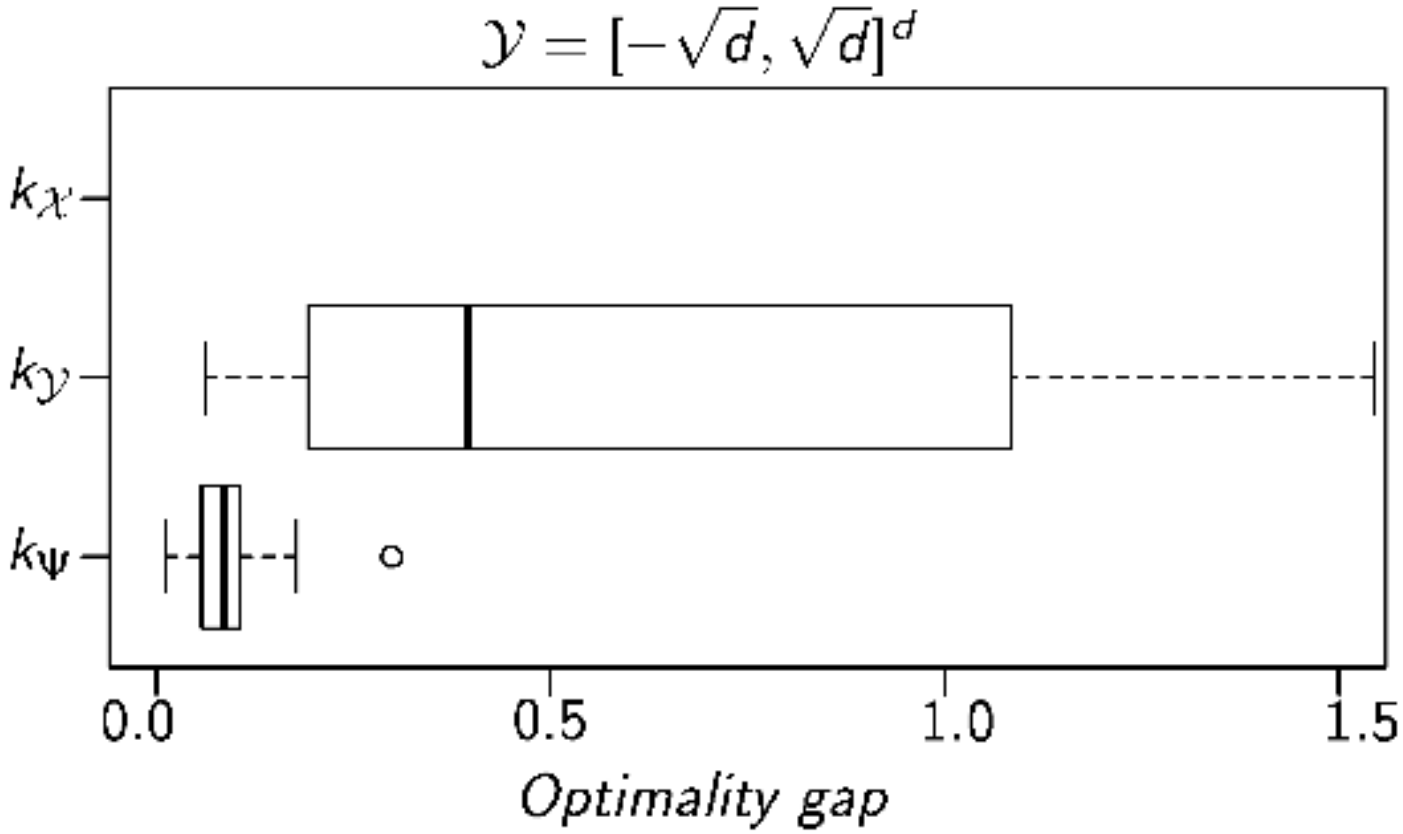
250 evaluations, $d = 6$, $D = 25$, 50 repetitions:



k_ψ is again the more robust for one embedding.

Results on the Hartman6 test function

250 evaluations, $d = 6$, $D = \mathbf{1000}$, 50 repetitions:



k_Ψ is again the more robust for one embedding.

Outline

Concepts in Multi-objective Optimization (MOO)

A solution minimizing every objective at once usually does not exist.

Concepts in Multi-objective Optimization (MOO)

A solution minimizing every objective at once usually does not exist.

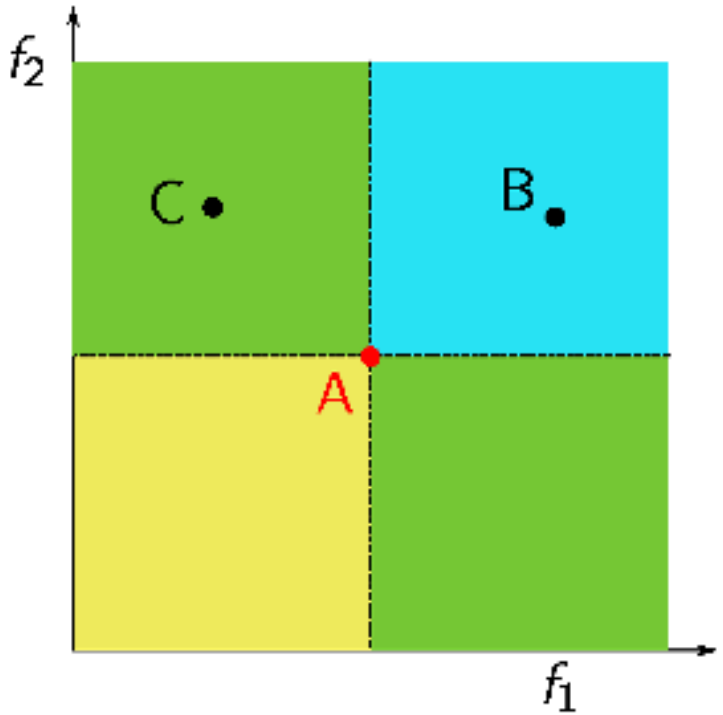
Pareto dominance

Vector A dominates vector B if:

- $\forall i \in \{1, \ldots, n\},\ a_i \leq b_i$
- $\exists j \in \{1, \ldots, n\}, a_j < b_j$

Pareto optimality

A is Pareto optimal if it is non-dominated.



Concepts in Multi-objective Optimization (MOO)

A solution minimizing every objective at once usually does not exist.

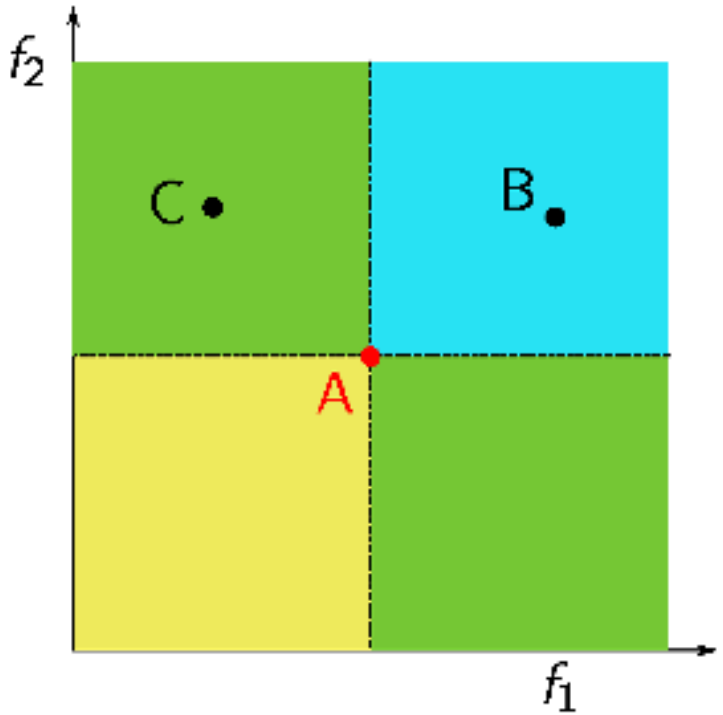
Pareto dominance

Vector A dominates vector B if:

- $\forall i \in \{1, \ldots, n\},\ a_i \leq b_i$
- $\exists j \in \{1, \ldots, n\}, a_j < b_j$

Pareto optimality

A is Pareto optimal if it is non-dominated.



- **Pareto set**: set of all optimal points in the variable space

Concepts in Multi-objective Optimization (MOO)

A solution minimizing every objective at once usually does not exist.

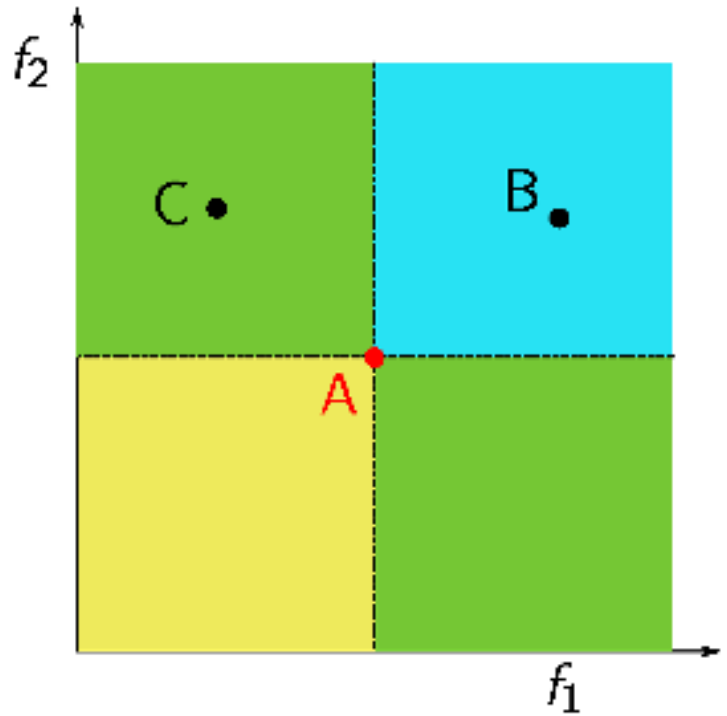
Pareto dominance

Vector A dominates vector B if:

- $\forall i \in \{1, \ldots, n\},\ a_i \leq b_i$
- $\exists j \in \{1, \ldots, n\}, a_j < b_j$

Pareto optimality

A is Pareto optimal if it is non-dominated.



- **Pareto set**: set of all optimal points in the variable space
- **Pareto front**: image of the Pareto set in the objective space

Concepts in Multi-objective Optimization (MOO)

A solution minimizing every objective at once usually does not exist.

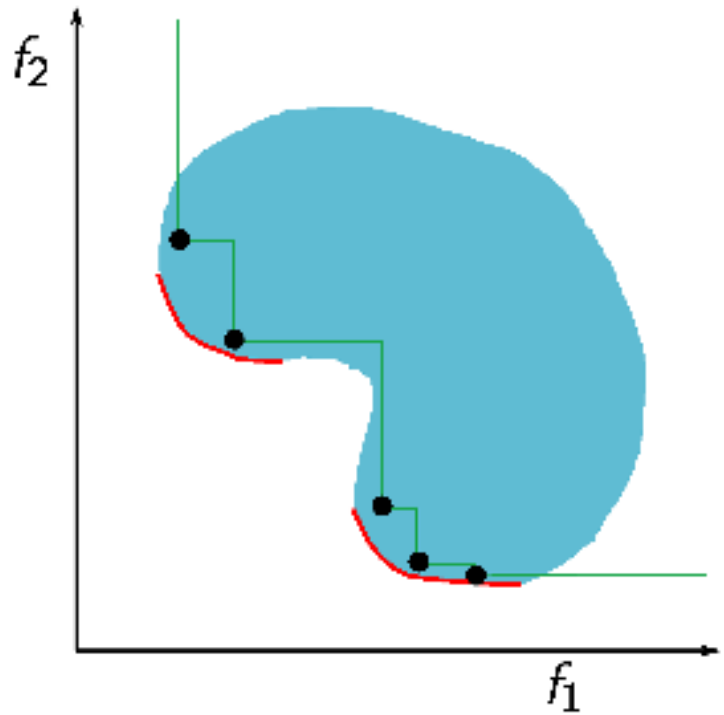
Pareto dominance

Vector A dominates vector B if:

- $\forall i \in \{1, \ldots, n\},\ a_i \leq b_i$
- $\exists j \in \{1, \ldots, n\}, a_j < b_j$

Pareto optimality

A is Pareto optimal if it is non-dominated.



- **Pareto set**: set of all optimal points in the variable space
- **Pareto front**: image of the Pareto set in the objective space
- Optimizers return non-dominated points close to the Pareto front

Extension to multi-objective (or constrained) optimization

Lemma

Assume $f = (f_1, \ldots, f_m)$, *where functions* $f_i : \mathbb{R}^D \mapsto \mathbb{R}, (1 \leq i \leq m)$ *have effective dimensionalities* d_{e_i} *(effective subspace* τ_i*) and a random matrix* $\mathbf{A} \in \mathbb{R}^{D \times d}$ *with independent entries sampled according to* $\mathcal{N}(0, 1)$, $d \geq Rank\left(\bigoplus_{i=1}^{m} \tau_i\right)$. *Then, with probability 1, for any* $\boldsymbol{x} \in \mathbb{R}^D$, $\exists \boldsymbol{y} \in \mathbb{R}^d$ *such that* $f(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{y})$.

Extension to multi-objective (or constrained) optimization

Lemma

Assume $f = (f_1, \ldots, f_m)$, *where functions* $f_i : \mathbb{R}^D \mapsto \mathbb{R}, (1 \leq i \leq m)$ *have effective dimensionalities* d_{e_i} *(effective subspace* τ_i*) and a random matrix* $\mathbf{A} \in \mathbb{R}^{D \times d}$ *with independent entries sampled according to* $\mathcal{N}(0, 1)$, $d \geq Rank\left(\bigoplus_{i=1}^{m} \tau_i\right)$. *Then, with probability 1, for any* $\boldsymbol{x} \in \mathbb{R}^D$, $\exists \boldsymbol{y} \in \mathbb{R}^d$ *such that* $f(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{y})$.

Sketch of proof

Extension of Theorem 2 from [WZM13] , considering the global effective subspace $\mathcal{T} \subset \mathbb{R}^D$.
First step: exhibit a $\mathbf{y}$ such that $f(\mathbf{x}) = f(\mathbf{x}_T + \mathbf{x}_\perp) = f(\mathbf{y})$.
Second step: using that with probability 1 the matrix $\mathbf{A}$ has rank d.

Extension to multi-objective (or constrained) optimization

Lemma

Assume $f = (f_1, \ldots, f_m)$, *where functions* $f_i : \mathbb{R}^D \mapsto \mathbb{R}, (1 \leq i \leq m)$ *have effective dimensionalities* d_{e_i} *(effective subspace* τ_i*) and a random matrix* $\mathbf{A} \in \mathbb{R}^{D \times d}$ *with independent entries sampled according to* $\mathcal{N}(0, 1)$, $d \geq Rank\left(\bigoplus_{i=1}^{m} \tau_i\right)$. *Then, with probability 1, for any* $\boldsymbol{x} \in \mathbb{R}^D$, $\exists \boldsymbol{y} \in \mathbb{R}^d$ *such that* $f(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{y})$.

Remark: when $\mathcal{X}$ is bounded, this is still true if the τ_i are aligned with the canonical basis.

Extension to multi-objective (or constrained) optimization

Lemma

Assume $f = (f_1, \ldots, f_m)$, *where functions* $f_i : \mathbb{R}^D \mapsto \mathbb{R}, (1 \leq i \leq m)$ *have effective dimensionalities* d_{e_i} *(effective subspace* τ_i*) and a random matrix* $\mathbf{A} \in \mathbb{R}^{D \times d}$ *with independent entries sampled according to* $\mathcal{N}(0, 1)$, $d \geq Rank\left(\bigoplus_{i=1}^{m} \tau_i\right)$. *Then, with probability 1, for any* $\boldsymbol{x} \in \mathbb{R}^D$, $\exists \boldsymbol{y} \in \mathbb{R}^d$ *such that* $f(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{y})$.

Remark: when $\mathcal{X}$ is bounded, this is still true if the τ_i are aligned with the canonical basis.

Problem: compared to the mono-objective case, the risk of missing parts of the Pareto front is even higher.

Extension to multi-objective (or constrained) optimization

Lemma

Assume $f = (f_1, \ldots, f_m)$, *where functions* $f_i : \mathbb{R}^D \mapsto \mathbb{R}, (1 \leq i \leq m)$ *have effective dimensionalities* d_{e_i} *(effective subspace* τ_i*) and a random matrix* $\mathbf{A} \in \mathbb{R}^{D \times d}$ *with independent entries sampled according to* $\mathcal{N}(0, 1)$, $d \geq Rank\left(\bigoplus_{i=1}^{m} \tau_i\right)$. *Then, with probability 1, for any* $\boldsymbol{x} \in \mathbb{R}^D$, $\exists \boldsymbol{y} \in \mathbb{R}^d$ *such that* $f(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{y})$.

Remark: when $\mathcal{X}$ is bounded, this is still true if the τ_i are aligned with the canonical basis.

Problem: compared to the mono-objective case, the risk of missing parts of the Pareto front is even higher.

Difference: use of a multi-objective version of the EI, e.g. the Expected Hypervolume Improvement [EDK11]

Selecting bounds for $\mathcal{Y}$ (cont'd)

Problem: ensuring that an optimum/the Pareto set is contained in $\mathcal{Y}$

Minimal domain $\mathcal{Z}$

The minimal star domain $\mathcal{Z} \subset \mathbb{R}^d$ with a one to one correspondence with $p_{\mathcal{X}}(\mathbf{A}\mathbb{R}^d)$ is: $\mathcal{Z} = \left\{\mathbf{y} \in \mathbb{R}^d, \forall \mathbf{x} \in p_{\mathcal{X}}(\mathbf{A}\mathbb{R}^d), \exists! \mathbf{y} \text{ s.t. } \mathbf{x} = p_{\mathcal{X}}(\mathbf{A}\mathbf{y})\right\}$

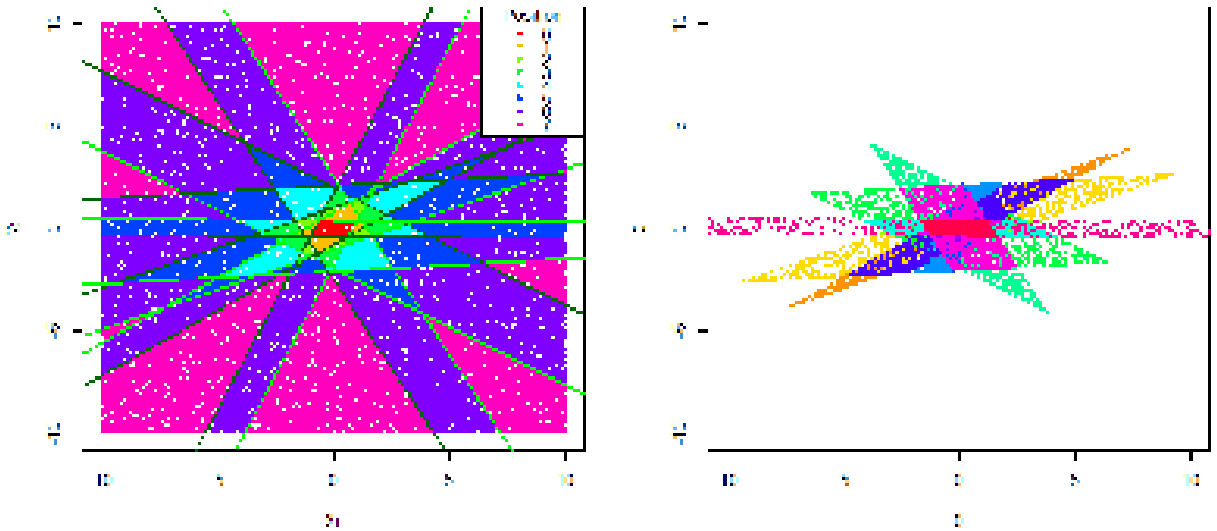
Selecting bounds for $\mathcal{Y}$ (cont'd)

Problem: ensuring that an optimum/the Pareto set is contained in $\mathcal{Y}$

Minimal domain $\mathcal{Z}$

The minimal star domain $\mathcal{Z} \subset \mathbb{R}^d$ with a one to one correspondence with $p_{\mathcal{X}}(\mathbf{A}\mathbb{R}^d)$ is: $\mathcal{Z} = \left\{\mathbf{y} \in \mathbb{R}^d, \forall \mathbf{x} \in p_{\mathcal{X}}(\mathbf{A}\mathbb{R}^d), \exists! \mathbf{y} \text{ s.t. } \mathbf{x} = p_{\mathcal{X}}(\mathbf{A}\mathbf{y})\right\}$

$$\mathcal{Z} = \bigcup_{1 \leq i_1 < \cdots < i_d \leq D} (H_{i_1} \cap \cdots \cap H_{i_d}) \text{ with } H_i = \left\{\mathbf{y} \in \mathbb{R}^d, -1 \leq A_i \mathbf{y} \leq 1\right\}$$



Selecting bounds for $\mathcal{Y}$ (cont'd)

In practice, $\mathcal{Y}$ is an hypercube: $\mathcal{Y} = [-\gamma, \gamma]^d$.
$\rightarrow$ How to select γ s.t. $\mathcal{Z} \subset \mathcal{Y}$?

Selecting bounds for $\mathcal{Y}$ (cont'd)

In practice, $\mathcal{Y}$ is an hypercube: $\mathcal{Y} = [-\gamma, \gamma]^d$.
$\rightarrow$ How to select γ s.t. $\mathcal{Z} \subset \mathcal{Y}$?

One can show that $\arg\max_{\mathbf{z} \in \mathcal{Z}} \|\mathbf{z}\|_d$ is a vertice of $\mathcal{X}$.

But: combinatorially intractable to find it.

Selecting bounds for $\mathcal{Y}$ (cont'd)

In practice, $\mathcal{Y}$ is an hypercube: $\mathcal{Y} = [-\gamma, \gamma]^d$.
$\rightarrow$ How to select γ s.t. $\mathcal{Z} \subset \mathcal{Y}$?

One can show that $\arg\max_{\mathbf{z} \in \mathcal{Z}} \|\mathbf{z}\|_d$ is a vertice of $\mathcal{X}$.

But: combinatorially intractable to find it.

Proposition (case $d = 2$)

If rows of $\mathbf{A}$, $\mathbf{A}_i$, $1 \leq i \leq D$, are vertices of a convex regular polygon in $\mathbb{R}^2$, then it is sufficient to take
$\gamma = \max\left(\left\|\mathbf{A}_{1,2}^{-1}[1,1]^T\right\|_d, \left\|\mathbf{A}_{1,2}^{-1}[1,-1]^T\right\|_d\right)$.
In addition, it minimizes the proportion of redundancy in $\mathcal{Y}$.

Selecting bounds for $\mathcal{Y}$ (cont'd)

In practice, $\mathcal{Y}$ is an hypercube: $\mathcal{Y} = [-\gamma, \gamma]^d$.
$\rightarrow$ How to select γ s.t. $\mathcal{Z} \subset \mathcal{Y}$?

One can show that $\arg\max_{\mathbf{z} \in \mathcal{Z}} \|\mathbf{z}\|_d$ is a vertice of $\mathcal{X}$.

But: combinatorially intractable to find it.

Proposition (case $d = 2$)

If rows of $\mathbf{A}$, $\mathbf{A}_i$, $1 \leq i \leq D$, are vertices of a convex regular polygon in $\mathbb{R}^2$, then it is sufficient to take
$\gamma = \max\left(\left\|\mathbf{A}_{1,2}^{-1}[1,1]^T\right\|_d, \left\|\mathbf{A}_{1,2}^{-1}[1,-1]^T\right\|_d\right)$.
In addition, it minimizes the proportion of redundancy in $\mathcal{Y}$.

It does not generalize with higher d: there exists no regular convex polytope for arbitrary D.

Approximating a regular polytope

Denote $\mathcal{P}$: $\left\{\mathbf{y} \in \mathbb{R}^d, 1 \leq i \leq D, -1 \leq \mathbf{A}_i\mathbf{y} \leq 1\right\}$

We want to modify $\mathbf{A}$ to approximate a regular convex polytope $\mathcal{P}$.

Approximating a regular polytope

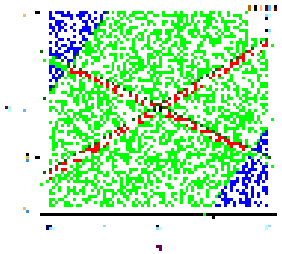
Denote $\mathcal{P}$: $\left\{\mathbf{y} \in \mathbb{R}^d, 1 \leq i \leq D, -1 \leq \mathbf{A}_i\mathbf{y} \leq 1\right\}$

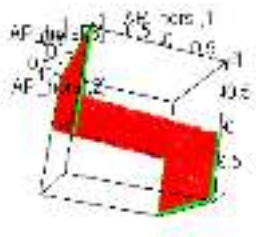
We want to modify $\mathbf{A}$ to approximate a regular convex polytope $\mathcal{P}$.

The matrix $\widetilde{\mathbf{A}}$ is obtained starting from $\mathbf{A}$ in two steps:

1. normalizing the rows of $\mathbf{A}$, to obtain points on the unit d-sphere
2. spreading of points on the unit d-sphere, using a potential:

$$\max_{\mathbf{x}_1,\dots\mathbf{x}_D\in\mathbf{S}} \min_{1\leq j<k\leq D} \|\mathbf{x}_j - \mathbf{x}_k\|_d$$





Initial matrix $\mathbf{A}$

Approximating a regular polytope

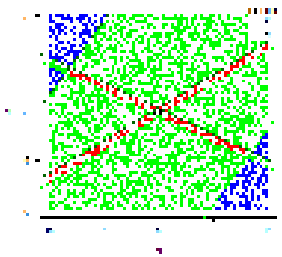
Denote $\mathcal{P}$: $\left\{\mathbf{y} \in \mathbb{R}^d, 1 \leq i \leq D, -1 \leq \mathbf{A}_i\mathbf{y} \leq 1\right\}$

We want to modify $\mathbf{A}$ to approximate a regular convex polytope $\mathcal{P}$.

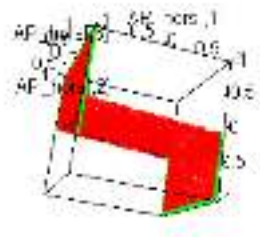
The matrix $\widetilde{\mathbf{A}}$ is obtained starting from $\mathbf{A}$ in two steps:

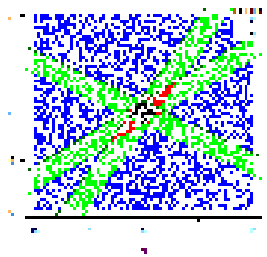
1. normalizing the rows of $\mathbf{A}$, to obtain points on the unit d-sphere
2. spreading of points on the unit d-sphere, using a potential:

$$\max_{\mathbf{x}_1,\ldots\mathbf{x}_D\in\mathbf{S}} \min_{1\leq j<k\leq D} \|\mathbf{x}_j - \mathbf{x}_k\|_d$$

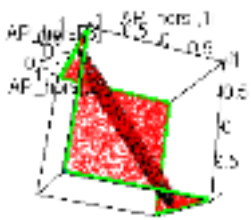


Initial matrix $\mathbf{A}$





Normalization



Approximating a regular polytope

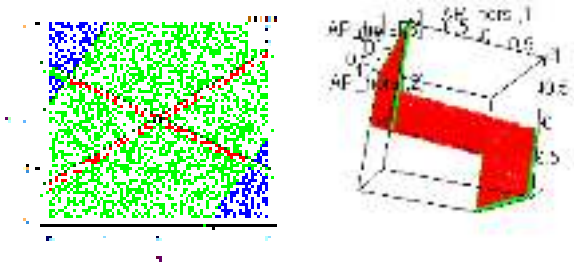
Denote $\mathcal{P}$: $\left\{\mathbf{y} \in \mathbb{R}^d, 1 \leq i \leq D, -1 \leq \mathbf{A}_i\mathbf{y} \leq 1\right\}$

We want to modify $\mathbf{A}$ to approximate a regular convex polytope $\mathcal{P}$.

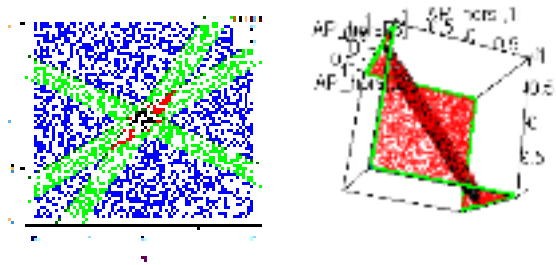
The matrix $\widetilde{\mathbf{A}}$ is obtained starting from $\mathbf{A}$ in two steps:

1. normalizing the rows of $\mathbf{A}$, to obtain points on the unit d-sphere
2. spreading of points on the unit d-sphere, using a potential:

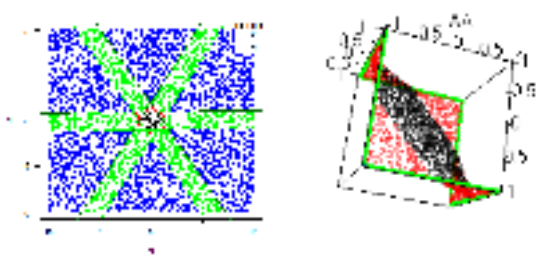
$$\max_{\mathbf{x}_1,\ldots\mathbf{x}_D \in \mathbf{S}} \min_{1 \leq j < k \leq D} \|\mathbf{x}_j - \mathbf{x}_k\|_d$$



Initial matrix $\mathbf{A}$



Normalization



$\widetilde{\mathbf{A}}$

Additional properties of the approximation

Proposition

After normalization of the rows of $\mathbf{A}$:

- the unit sphere in $\mathbb{R}^d$ is the biggest sphere enclosing only points mapping onto $\mathbf{A}\mathbb{R}^d \cap \mathcal{X}$
- the number of vertices of the polytope $\mathcal{P}$ is maximal.

$\rightarrow$ this gives a lower bound for $\mathcal{Y}$: $[-1, 1]^d$

Additional properties of the approximation

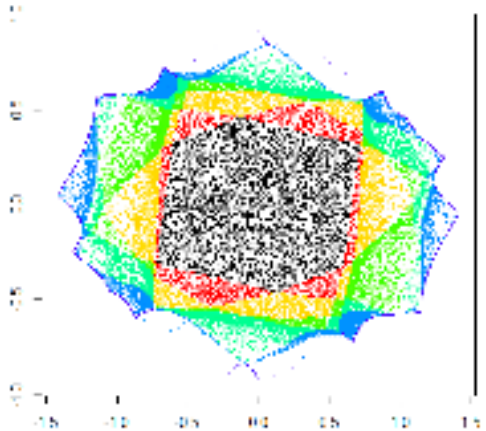
Proposition

After normalization of the rows of $\mathbf{A}$:

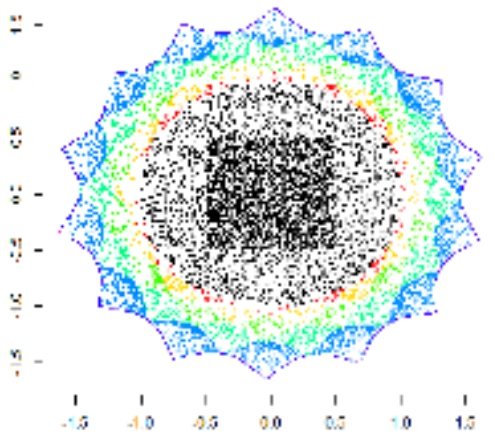
- the unit sphere in $\mathbb{R}^d$ is the biggest sphere enclosing only points mapping onto $\mathbf{A}\mathbb{R}^d \cap \mathcal{X}$
- the number of vertices of the polytope $\mathcal{P}$ is maximal.

$\rightarrow$ this gives a lower bound for $\mathcal{Y}$: $[-1, 1]^d$

By-product: less deformation in the warped space $\mathbf{A}^{-1}\Psi(\mathcal{Y})$



Random $\mathbf{A}$



$\tilde{\mathbf{A}}$

Additional properties of the approximation

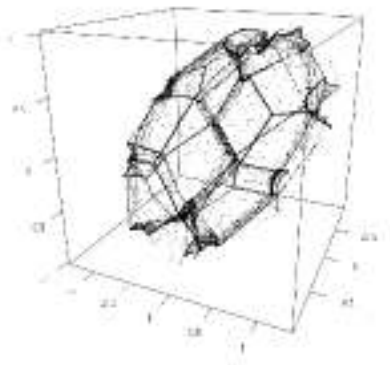
Proposition

After normalization of the rows of $\mathbf{A}$:

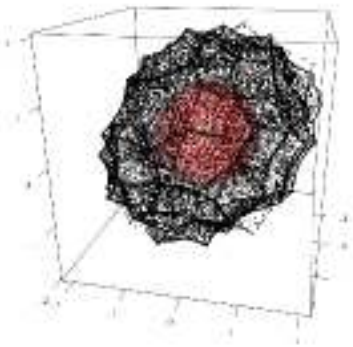
- the unit sphere in $\mathbb{R}^d$ is the biggest sphere enclosing only points mapping onto $\mathbf{A}\mathbb{R}^d \cap \mathcal{X}$
- the number of vertices of the polytope $\mathcal{P}$ is maximal.

$\rightarrow$ this gives a lower bound for $\mathcal{Y}$: $[-1, 1]^d$

By-product: less deformation in the warped space $\mathbf{A}^{-1}\Psi(\mathcal{Y})$



Random $\mathbf{A}$

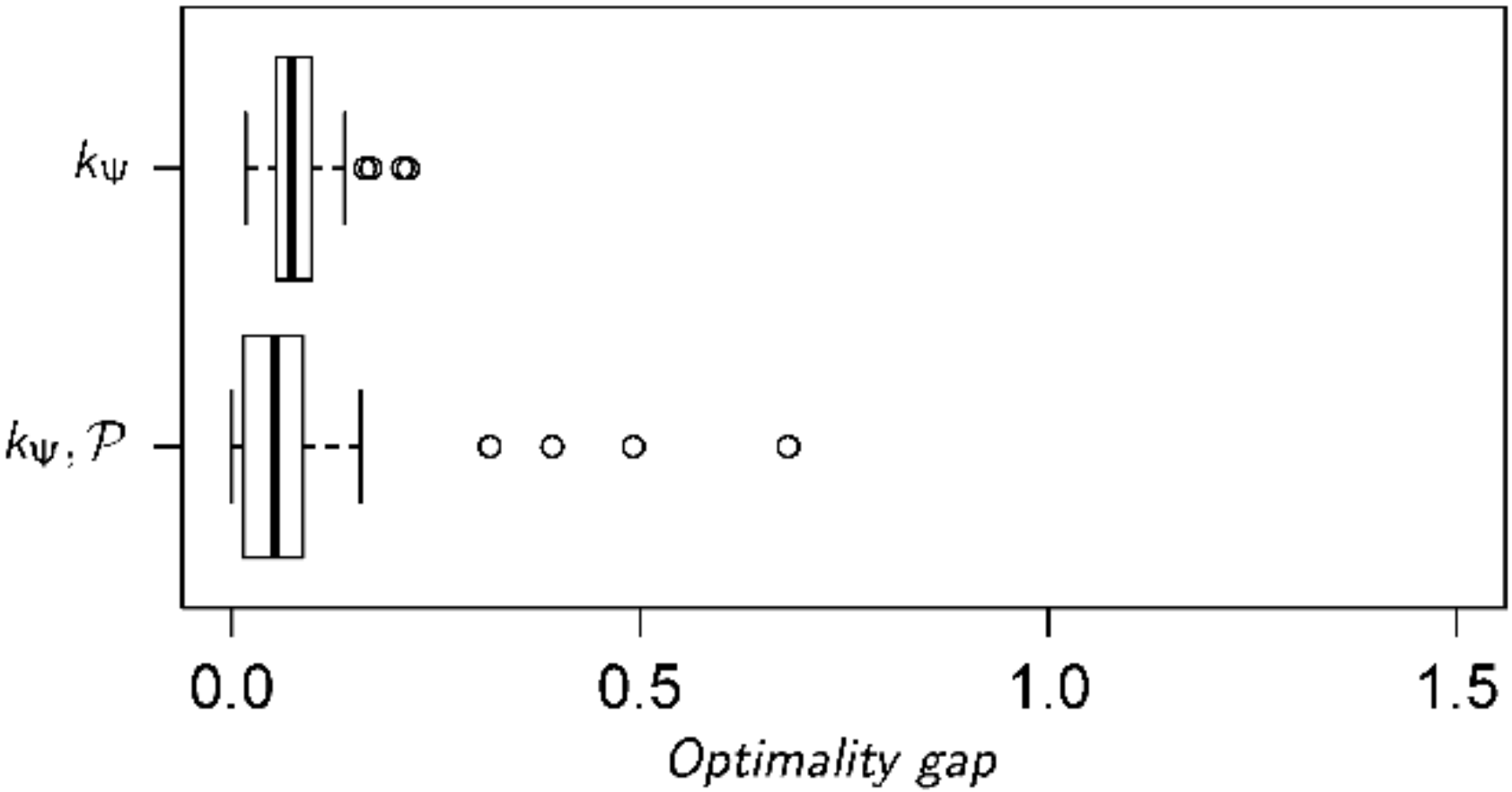


$\widetilde{\mathbf{A}}$

Experimental validation

Mono-objective

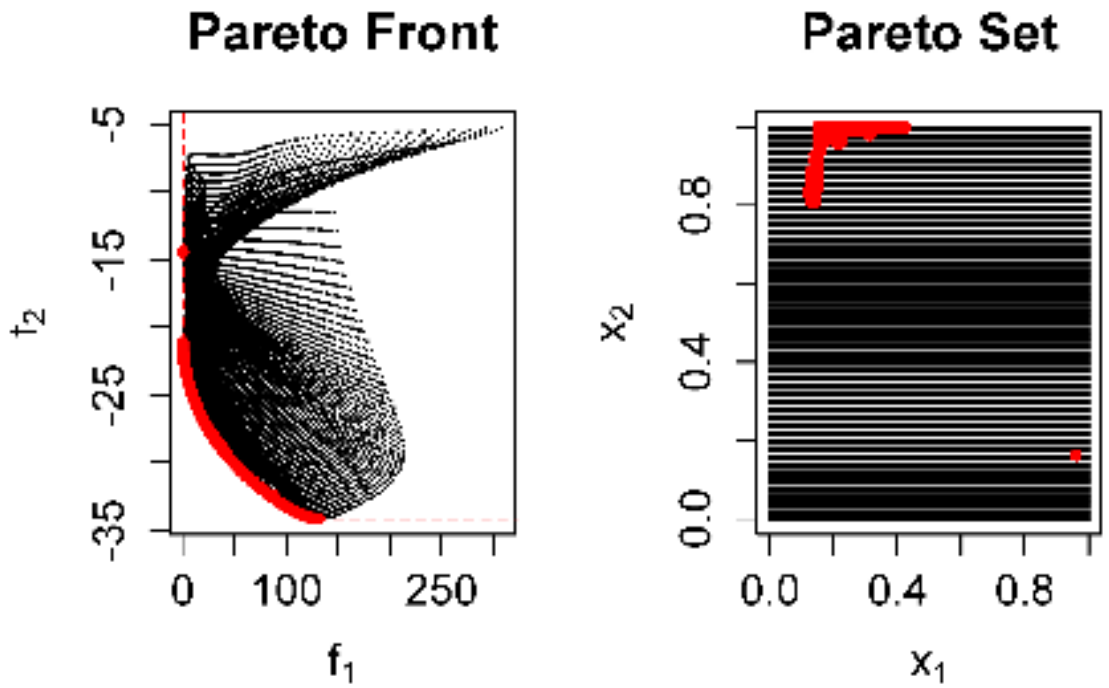
Hartman6 test function, 250 evaluations, $d = 6$, $D = 25$, 50 repetitions:



Experimental validation

Multi-objective

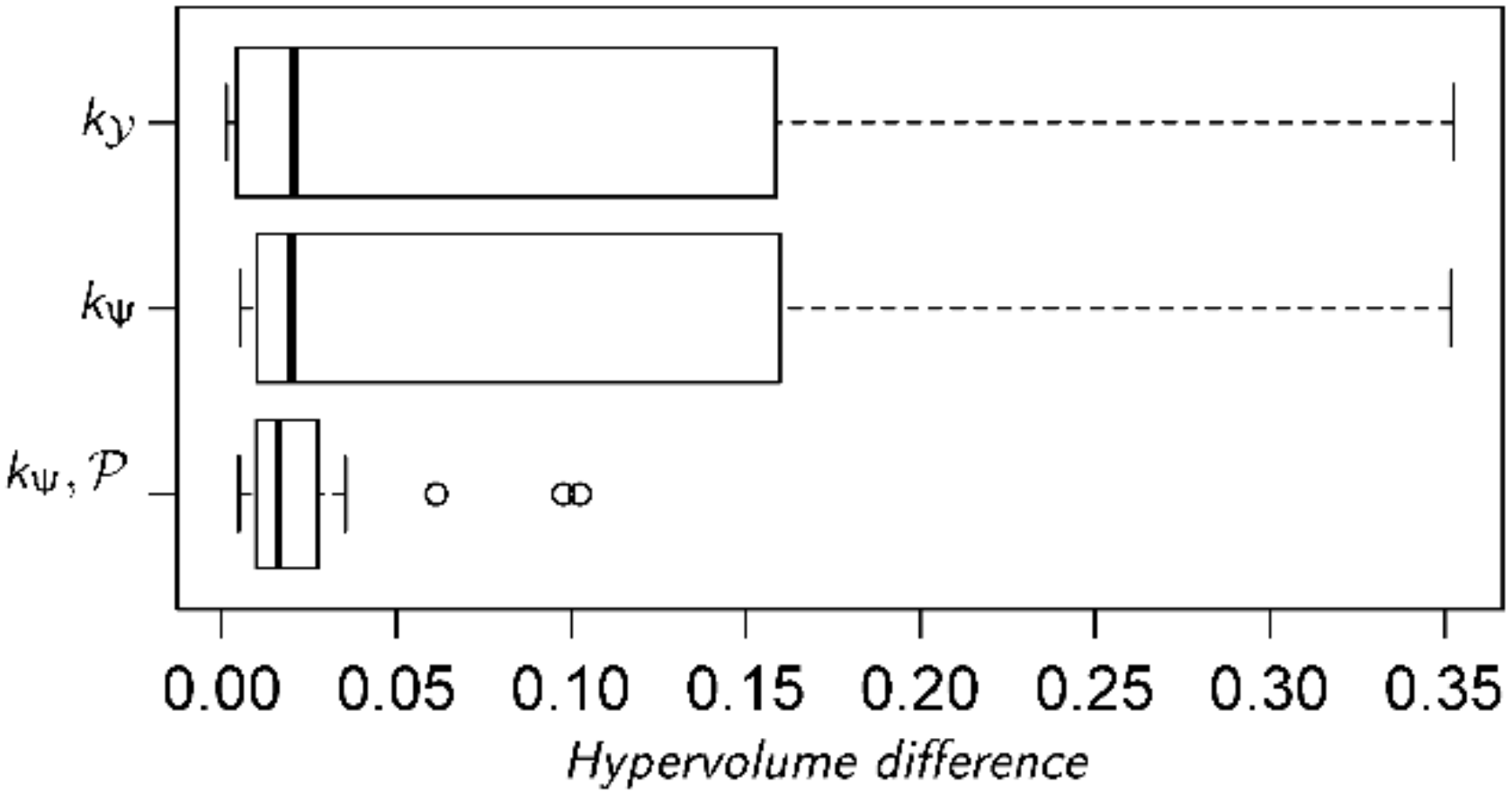
First test function, 100 evaluations, $d = 2$, $D = 25$, 25 repetitions:



Experimental validation

Multi-objective

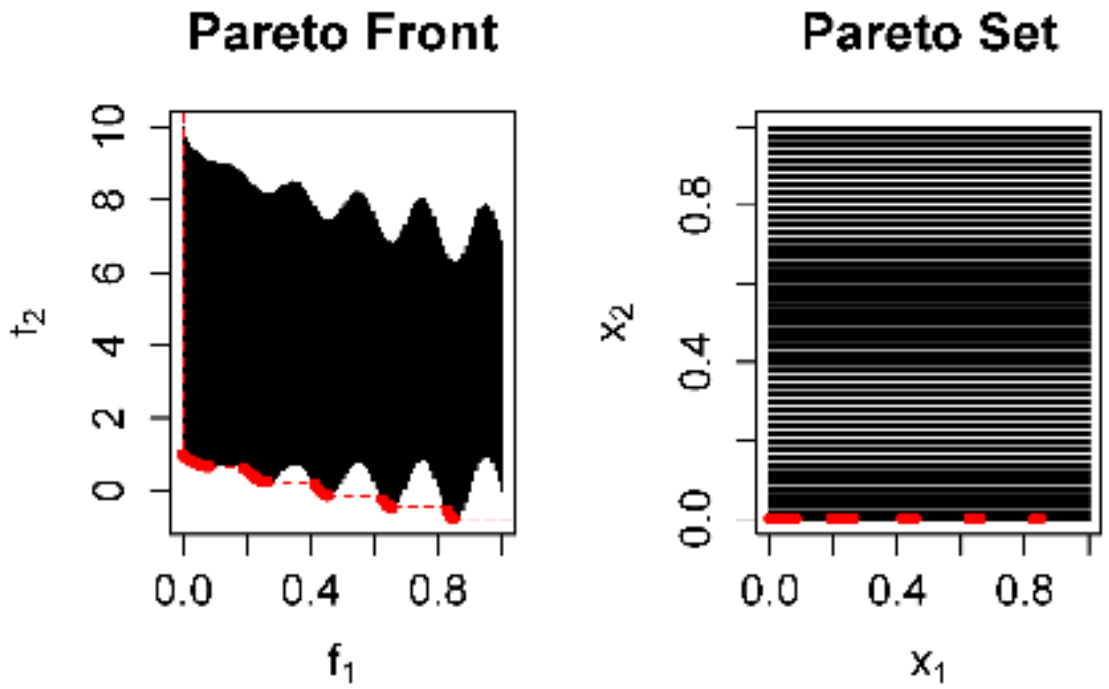
First test function, 100 evaluations, $d = 2$, $D = 25$, 25 repetitions:



Experimental validation

Multi-objective

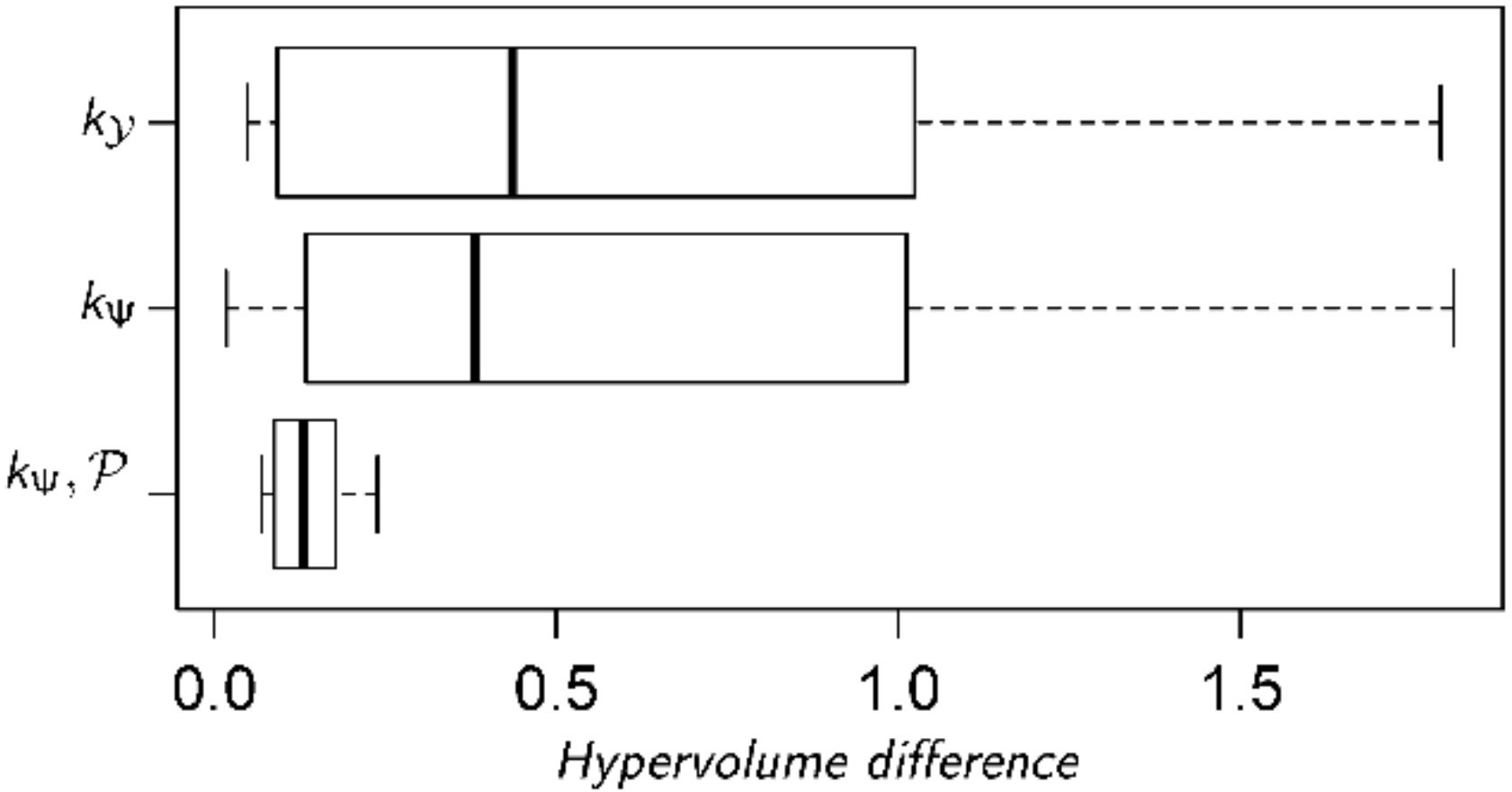
ZDT3 test function, 100 evaluations, $d = 2$, $D = 25$, 25 repetitions:



Experimental validation

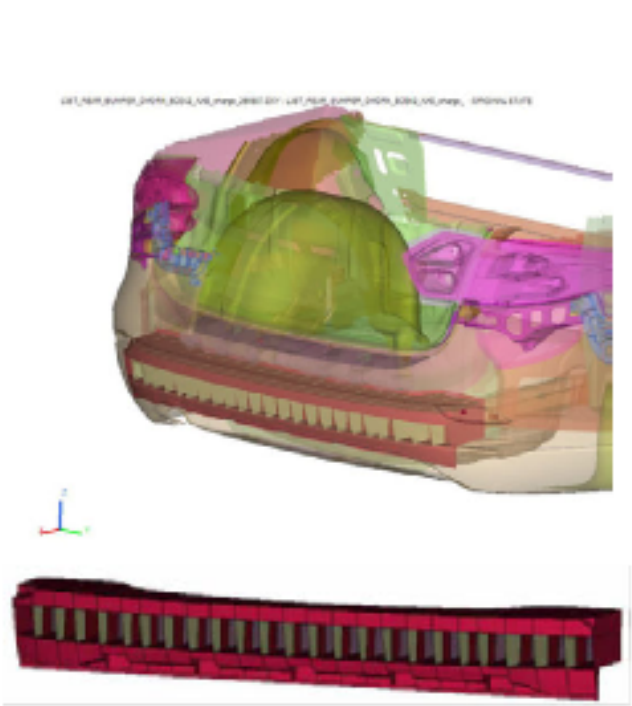
Multi-objective

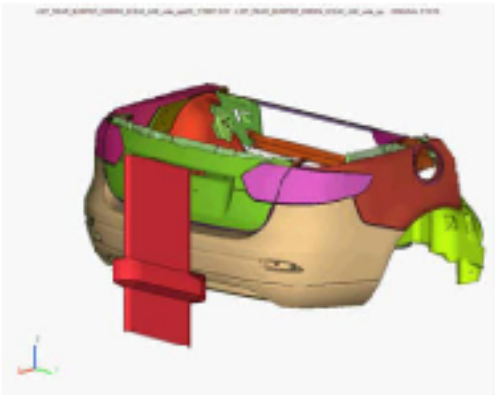
ZDT3 test function, 100 evaluations, $d = 2$, $D = 25$, 25 repetitions:

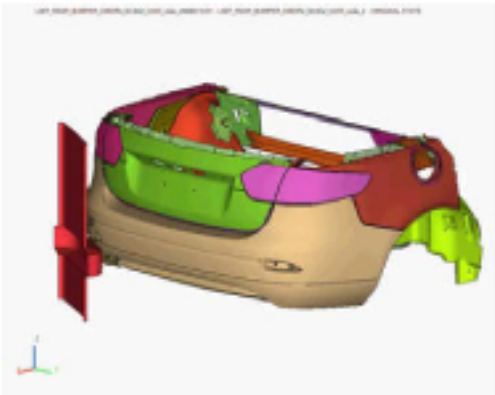


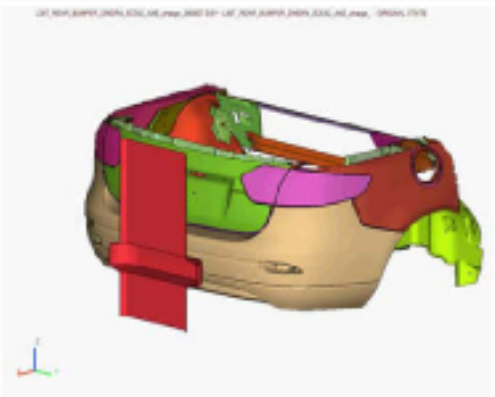
Outline

Test case: rear shock absorber L38 - ECE42







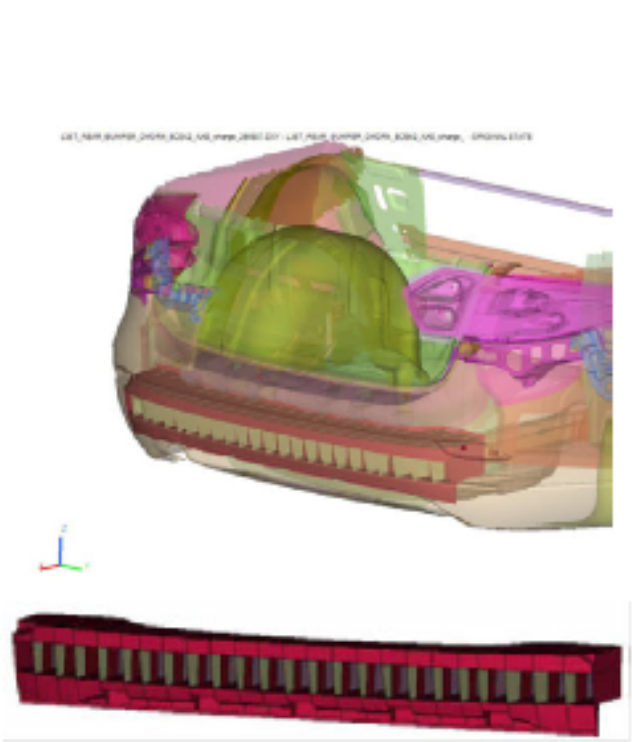


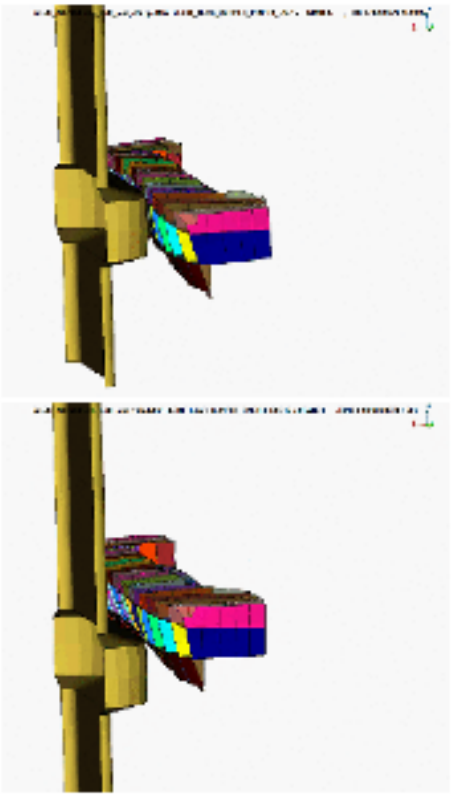
Axial impact: 4km/h, 80mm of intrusion max

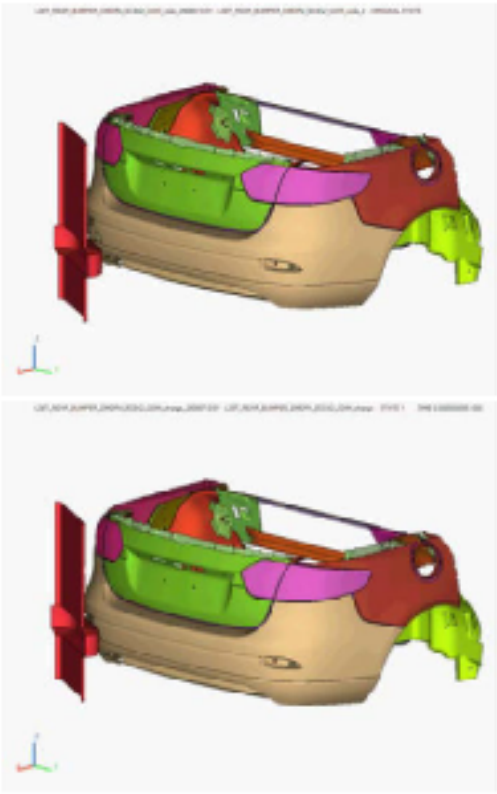
Lateral impact: 2.5km/h, 55mm of intrusion max

47 parameters (thicknesses)

Test case: rear shock absorber L38 - ECE42

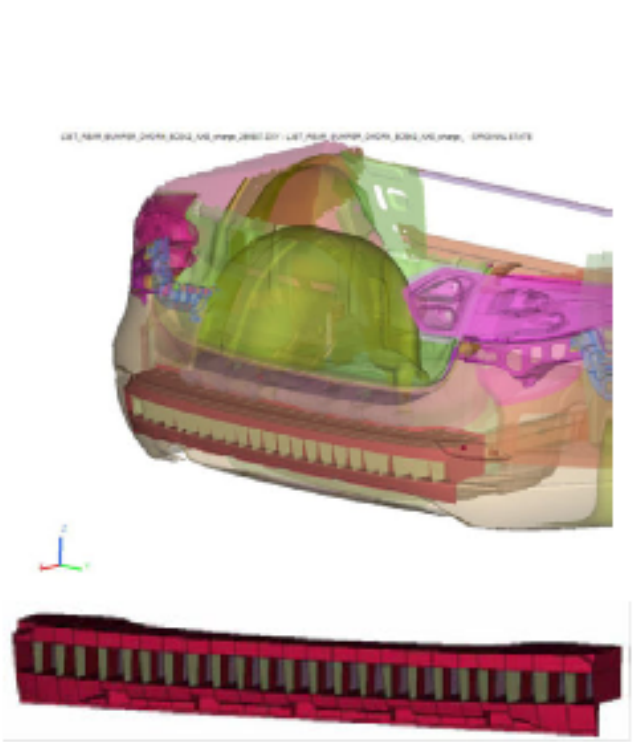


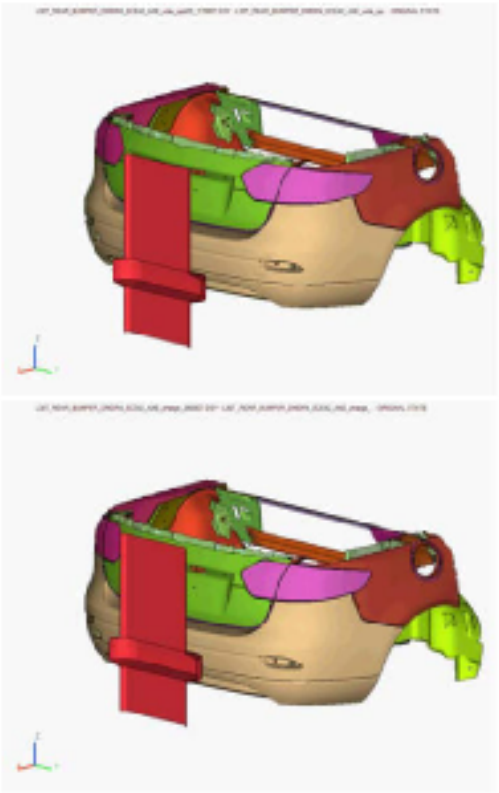


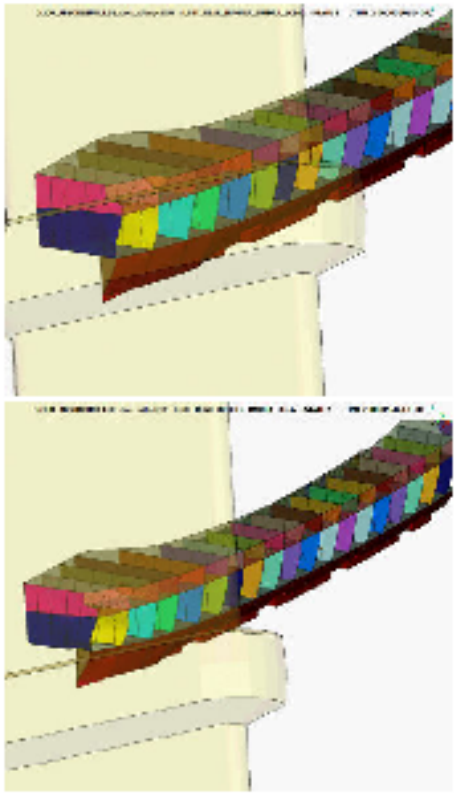


Axial impact: 4km/h, 80mm of intrusion max
Lateral impact: 2.5km/h, 55mm of intrusion max
47 parameters (thicknesses)

Test case: rear shock absorber L38 - ECE42







Axial impact: 4km/h, 80mm of intrusion max
Lateral impact: 2.5km/h, 55mm of intrusion max
47 parameters (thicknesses)

Results

	Mass (g)	Axial Loaded	Axial Empty	Lateral Loaded	Lateral Empty
Specifications		$<$ 80mm	$<$80mm	$<$ 55mm	$<$ 55mm
Reference	3375	74.1	72.9	25	35
Older study (2007)	2942 (-12.8%)	78.4	77.6	26	33.4
Constrained BO	2762 (-18.2%)	80.0	78.7	34.4	50.9
Constrained REMBO	2904 (-14,0%)	79.0	80.0	27.1	53.3

# of simulations	Axial Loaded	Axial Empty	Lateral Loaded	Lateral Empty
Older study (2007)	286	286	62	92
Constrained BO	210	210	177	142
Constrained REMBO	140	140	140	140

Details for REMBO:

- $d = 18$
- influential subspace is **not** aligned with the canonical basis of $\mathcal{X}$.

Outline

Conclusion

Contributions:

- extension to constrained/multi-objective optimization

Conclusion

Contributions:

- extension to constrained/multi-objective optimization
- design of a specialized kernel for REMBO which retains some high-dimensional correlation information in a low dimensional space

Conclusion

Contributions:

- extension to constrained/multi-objective optimization
- design of a specialized kernel for REMBO which retains some high-dimensional correlation information in a low dimensional space
- customized tuning of the parameter space through modifications of $\mathbf{A}$ to improve bounds for $\mathcal{Y}$

Conclusion

Contributions:

- extension to constrained/multi-objective optimization
- design of a specialized kernel for REMBO which retains some high-dimensional correlation information in a low dimensional space
- customized tuning of the parameter space through modifications of $\mathbf{A}$ to improve bounds for $\mathcal{Y}$

Conclusion

Contributions:

- extension to constrained/multi-objective optimization
- design of a specialized kernel for REMBO which retains some high-dimensional correlation information in a low dimensional space
- customized tuning of the parameter space through modifications of $\mathbf{A}$ to improve bounds for $\mathcal{Y}$

Result: significant improvement of the performance of REMBO with a single embedding

Conclusion

Contributions:

- extension to constrained/multi-objective optimization
- design of a specialized kernel for REMBO which retains some high-dimensional correlation information in a low dimensional space
- customized tuning of the parameter space through modifications of $\mathbf{A}$ to improve bounds for $\mathcal{Y}$

Result: significant improvement of the performance of REMBO with a single embedding

Ongoing work:

- study of other optimality criteria for $\mathbf{A}$

Conclusion

Contributions:

- extension to constrained/multi-objective optimization
- design of a specialized kernel for REMBO which retains some high-dimensional correlation information in a low dimensional space
- customized tuning of the parameter space through modifications of $\mathbf{A}$ to improve bounds for $\mathcal{Y}$

Result: significant improvement of the performance of REMBO with a single embedding

Ongoing work:

- study of other optimality criteria for $\mathbf{A}$
- further specialization of kernels to account for the different cases: $\mathcal{T}$ aligned with $\mathcal{X}$, $\mathcal{T} \not\perp \mathbf{A}\mathcal{Y}$, $\mathcal{T} \perp \mathbf{A}\mathcal{Y}$

Bibliography I

M. B., D. Ginsbourger, and O. Roustant

A warped kernel improving robustness in Bayesian optimization via random embeddings

In: Proceedings of the International Conference on Learning and Intelligent Optimization (LION'15) (To appear)

M. B., V. Picheny

GPareto: Gaussian Processes for Pareto Front Estimation and Optimization

R package version 1.0.0, 2014

M. Emmerich, A. Deutz, J. Klinkenberg

Hypervolume-based expected improvement: Monotonicity properties and exact computation

Evolutionary Computation (CEC), 2011

C. Rasmussen, C. Williams

Gaussian Processes for Machine Learning

MIT Press, 2006

Z. Wang, M. Zoghi, D. Matheson, F. Hutter, and N. de Freitas

Bayesian optimization in a billion dimensions via random embeddings

In IJCAI, 2013

D. Jones, M. Schonlau and W. Welch

Efficient global optimization of expensive black-box functions

Journal of Global Optimization, 1998

J. Snoek, K. Swersky, R. S. Zemel, S. Richard and R. P. Adams

Input Warping for Bayesian Optimization of Non-stationary Functions

In ICML, 2014

Bibliography II

O. Roustant, D. Ginsbourger and Y. Deville

DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization

Journal of Statistical Software, 2012

M. L. Stein

Interpolation of spatial data: some theory for kriging

Springer, 1999

G. Matheron

Principles of geostatistics

In Economic geology, 1963

Thank you for your attention!

Questions?