Algorithmic aspects of large scale optimal experimental design

Guillaume SAGNOL

Zuse Institut Berlin (ZIB)

February 8, 2012
Outline

1. Optimal Design of Experiments

2. Application to Network Monitoring

3. Algorithms for Optimal Experimental Design
   - Classic algorithms
   - Recent developments
   - New search directions
Outline

1. Optimal Design of Experiments
2. Application to Network Monitoring
3. Algorithms for Optimal Experimental Design
   - Classic algorithms
   - Recent developments
   - New search directions
We want to estimate the vector of parameters \( \theta = [\theta_1, \ldots, \theta_m]^T \).
We want to estimate the vector of parameters \( \theta = [\theta_1, \ldots, \theta_m]^T \).

We have \( s \) experiments available:

\[
\begin{align*}
\mathbf{y}_1 &= A_1 \theta + \epsilon_1 \\
\mathbf{y}_2 &= A_2 \theta + \epsilon_2 \\
\mathbf{y}_3 &= A_3 \theta + \epsilon_3 \\
& \vdots \\
\mathbf{y}_s &= A_s \theta + \epsilon_s
\end{align*}
\]
Optimal Design of Experiments

We want to estimate the vector of parameters 
\( \theta = [\theta_1, \ldots, \theta_m]^T \).

We have \( s \) experiments available:

\[
\begin{align*}
    y_1 &= A_1 \theta + \epsilon_1 \\
    y_2 &= A_2 \theta + \epsilon_2 \\
    y_3 &= A_3 \theta + \epsilon_3 \\
    \vdots \\
    y_s &= A_s \theta + \epsilon_s
\end{align*}
\]

**GOAL:**
1. Choose how many times to perform each experiment.
2. Continuous relaxation on the unit \( s \)-simplex: *distribute the experimental effort* \( w_i \) (with \( \sum_{i=1}^{s} w_i = 1 \)).
minimizing the variance of the best estimator

Under classical assumptions (e.g. noises have unit variance and are independent),

**Theorem [Gauss-Markov]**

For every linear unbiased estimator $\hat{\theta}$ of the unknown parameter $\theta$, we have:

$$\text{Var}[\hat{\theta}] \preceq \left( \sum_{i=1}^{s} w_i A_i^T A_i \right)^{-1}.$$  

Moreover this lower bound is attained by the estimator from least square theory.
Two common criterions

**Definition: Information matrix**

\[ M(w) := \sum_{i=1}^{s} w_i A_i^T A_i \] is the information matrix of \( w \).

**Common approach:** Minimize a scalar function of \( M(w)^{-1} \).
Two common criterions

Definition: Information matrix

\[ M(w) := \sum_{i=1}^{s} w_i A_i^T A_i \] is the information matrix of \( w \).

Common approach: Minimize a scalar function of \( M(w)^{-1} \).

Two standard criterions

- \( c \)-optimality: given \( c \in \mathbb{R}^m \),

\[
\min \left\{ c^T M(w)^{-1} c : \ w \in \mathbb{R}_+^s, \ \sum_i w_i = 1 \right\}
\]

relevant when we want to estimate \( c^T \theta \)
Two common criterions

**Definition: Information matrix**

\[ M(\mathbf{w}) := \sum_{i=1}^{s} w_i A_i^T A_i \] is the *information matrix* of \( \mathbf{w} \).

**Common approach:** Minimize a scalar function of \( M(\mathbf{w})^{-1} \).

**Two standard criterions**

- **\( c \)-optimality:** given \( \mathbf{c} \in \mathbb{R}^m \),

  \[
  \min \left\{ \mathbf{c}^T M(\mathbf{w})^{-1} \mathbf{c} : \quad \mathbf{w} \in \mathbb{R}_+^s, \; \sum_{i} w_i = 1 \right\}
  \]

- **\( A \)-optimality:**

  \[
  \min \left\{ \text{trace} \; M(\mathbf{w})^{-1} : \quad \mathbf{w} \in \mathbb{R}_+^s, \; \sum_{i} w_i = 1 \right\}
  \]
Given an eigenvalue decomposition

\[ M(w) = \sum \lambda_i u_i u_i^T, \]

the confidence ellipsoids of \( \hat{\theta} \) are centered in \( \theta \), and have semi-axis of length proportional to \( \frac{1}{\sqrt{\lambda_i}} \).
Geometric interpretation

Given an eigenvalue decomposition

\[ M(w) = \sum \lambda_i u_i u_i^T, \]

the confidence ellipsoids of \( \hat{\theta} \) are centered in \( \theta \), and have semi-axis of length proportional to \( \frac{1}{\sqrt{\lambda_i}} \).

\[ \Phi_A(w) \]

A–optimal design:

\[ \min \frac{1}{\lambda_1} + \frac{1}{\lambda_2} \]

minimize the diagonal of the bounding box.
Geometric interpretation

Given an eigenvalue decomposition

\[ M(\mathbf{w}) = \sum \lambda_i \mathbf{u}_i \mathbf{u}_i^T, \]

the confidence ellipsoids of \( \hat{\theta} \) are centered in \( \theta \), and have semi-axis of length proportional to \( \frac{1}{\sqrt{\lambda_i}} \).

\[ \Phi_c(\mathbf{w}) \]

\( c \)-optimal design:

minimize shadow along vector \( c \).
Geometric interpretation

Given an eigenvalue decomposition

\[ M(w) = \sum \lambda_i u_i u_i^T, \]

the confidence ellipsoids of \( \hat{\theta} \) are centered in \( \theta \), and have semi-axis of length proportional to \( \frac{1}{\sqrt{\lambda_i}} \).

\[ \Phi_D(w) \]

\[ \frac{1}{\sqrt{\lambda_2}} \]

\[ \frac{1}{\sqrt{\lambda_1}} \]

\[ u_2 \]

\[ u_1 \]

\[ \theta \]

D–optimal design:

\[ \max \lambda_1 \lambda_2 \]

minimize the volume.
Given an eigenvalue decomposition

\[ M(w) = \sum \lambda_i u_i u_i^T, \]

the confidence ellipsoids of \( \hat{\theta} \) are centered in \( \theta \), and have semi-axis of length proportional to \( \frac{1}{\sqrt{\lambda_i}} \).

\[ \Phi_E(w) \]

\( E \)-optimal design:

\[ \max \lambda_1 \]

minimize the largest semi-axis.
Given observation matrices $A_1, \ldots, A_s$, solve

$$\min_w \Phi \left( \sum_{i=1}^{s} w_i A_i^T A_i \right)$$

s. t. $w \geq 0$, $\sum_{i=1}^{s} w_i = 1$.

$\Phi$ is a function mapping $X \in \mathbb{S}_m^+$ to $\mathbb{R}$, that is nonincreasing with respect to Löwner ordering $\succeq$, e.g.

$$\Phi(X) = \text{trace } X^{-1} \quad (A-\text{optimality})$$

$$\Phi(X) = c^T X^{-1} c \quad (c-\text{optimality})$$
Outline

1. Optimal Design of Experiments
2. Application to Network Monitoring
3. Algorithms for Optimal Experimental Design
   - Classic algorithms
   - Recent developments
   - New search directions
We want to monitor certain quantities on a network, e.g.

- volume of the flows
- performance indicator
- nature of the traffic

**General Problem**
How many monitors should we place on each Node/Edge/OD Pair of the network?
We place a monitor on the blue edge:
⇒ we collect information about 4 OD flows:
example: OD-flows monitoring

We place a monitor on the blue edge:
⇒ we collect information about 4 OD flows:

- 1 → 4
- 1 → 5
- 2 → 4
- 2 → 5
example: OD-flows monitoring (continued)

In some applications, monitors can’t determine the source of the flows:

\[
A_{2\rightarrow 4} = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\end{pmatrix}
\]
### A Huge System
- Introduced in 2005
- Tax based on: distance and truck category (avg price: 0.17 Euros/km)
- Total revenue around 3 Billion Euros / year

### Control Forces
- Automatic: 300 toll checker gantries (Kontrollbrücke)
- Manual: 300 vehicles – 540 officiers from BAG (Federal ofOice of Freight)

### Complexity of BAG’s management
- Schedule of each officer
- Choose where and when to control
TODO: show NRW and Berlin, CPU time, what about whole Germany?
Outline

1. Optimal Design of Experiments
2. Application to Network Monitoring
3. Algorithms for Optimal Experimental Design
   - Classic algorithms
   - Recent developments
   - New search directions
Outline

1. Optimal Design of Experiments
2. Application to Network Monitoring
3. Algorithms for Optimal Experimental Design
   - Classic algorithms
   - Recent developments
   - New search directions
Wynn-Fedorov exchange algorithm

\[
\min \left\{ \text{trace } M(w)^{-1} : \ w \in \mathbb{R}_+^s, \ \sum_i w_i = 1, \ M(w) = \sum_i w_i A_i^T A_i \right\}
\]

Algorithm (Wynn 70, Fedorov 72)

This is a feasible descent algorithm:

- At step \( k \), find the atomic design \( e_{i_k} \) such that the directional derivative is maximum.

\[
i_k \leftarrow \arg\max_{i \in [s]} \| M(w^{(k)})^{-1} A_i \|_F
\]

- For a step of length \( \alpha_k \), move in the direction of \( e_{i_k} \):

\[
w^{(k+1)} \leftarrow (1 - \alpha_k) w^{(k)} + \alpha_k e_{i_k}.
\]
Titterington multiplicative algorithm

\[
\min \left\{ \text{trace } M(w)^{-1} : \quad w \in \mathbb{R}_+^s, \quad \sum_i w_i = 1, \quad M(w) = \sum_i w_i A_i^T A_i \right\}
\]

Algorithm (Titterington 76)

At step \( k \),

- For all \( i \in [s] \), compute the directional derivative
  \[
d_i \leftarrow \| M(w^{(k)})^{-1} A_i \|_F
\]

- For a parameter \( \lambda \), make the multiplicative updates:
  \[
  w_i^{(k+1)} \leftarrow w_i^{(k)} \frac{d_i^\lambda}{\Gamma},
  \]

where \( \Gamma \) is a normalizing constant.
Outline

1. Optimal Design of Experiments
2. Application to Network Monitoring
3. Algorithms for Optimal Experimental Design
   - Classic algorithms
   - Recent developments
   - New search directions
Semidefinite Programming

\[
\min \left\{ \text{trace } M(w)^{-1} : \quad w \in \mathbb{R}^s_+, \quad \sum_i w_i = 1, \quad M(w) = \sum_i w_i A_i^T A_i \right\}
\]

A-optimality SDP (Vandenberghe, Boyd, Wu 98)

\[
\begin{align*}
\min_{w, \ Y \in S_m} \quad & \text{trace } Y \\
\text{s. t.} \quad & \begin{pmatrix} M(w) & I \\ I & Y \end{pmatrix} \succeq 0, \\
& \sum_{i=1}^{s} w_i = 1, \quad \forall \ i \in [s], \ w_i \geq 0.
\end{align*}
\]
Elfving’s Theorem

In presence of scalar observations ($A_i = a_i^T$), we have:

**Theorem [Elfving,53]**

The design $w$ is $c$–optimal iff there exists $t > 0$ such that

$$tc = \sum_k w_k a_k \in \partial\left(\text{conv}\{\pm a_i, i = 1, \ldots, s\}\right).$$

$$\max_{\lambda, t} t \quad \text{s.t.} \quad tc = \sum_k \lambda_i a_i \quad \sum_k |\lambda_k| w_k \leq 1.$$
General case

**Theorem [DHL09], [Sag09,10]**

The design $\mathbf{w}$ is $\mathbf{c}$-optimal iff there exists $t > 0$ and unit vectors $\epsilon_k$ such that

$$t \mathbf{c} = \sum_k w_k A_k^T \epsilon_k \in \partial \left( \text{conv} \{ A_i^T \epsilon, \; i = 1\ldots,s, \; \| \epsilon \| \leq 1 \} \right).$$
An illustration in 3D

\[
\begin{align*}
A_1 &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix} \\
A_2 &= \begin{pmatrix} 0 & 0.5 & 2 \end{pmatrix} \\
A_3 &= \begin{pmatrix} 0 & -0.5 & 2 \end{pmatrix}
\end{align*}
\]
An illustration in 3D

\[ A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix} \]

\[ A_2 = \begin{pmatrix} 0 & 0.5 & 2 \end{pmatrix} \]

\[ A_3 = \begin{pmatrix} 0 & -0.5 & 2 \end{pmatrix} \]
$A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix}$

$A_2 = \begin{pmatrix} 0 & 0.5 & 2 \end{pmatrix}$

$A_3 = \begin{pmatrix} 0 & -0.5 & 2 \end{pmatrix}$
An illustration in 3D

The uppermost half of the generalized Elfving set

\[ A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix} \]

\[ A_2 = \begin{pmatrix} 0 & 0.5 & 2 \end{pmatrix} \]

\[ A_3 = \begin{pmatrix} 0 & -0.5 & 2 \end{pmatrix} \]
If we want to estimate \( (x + 2z) \) → Barycenter of \( a_{1(1)}, a_2 \) and \( a_3 \)
If we want to estimate \((x + y + z)\) → Barycenter of \(x_1\) and \(a_2\)

\[
A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix} \\
A_2 = \begin{pmatrix} 0 & 0.5 & 2 \end{pmatrix} \\
A_3 = \begin{pmatrix} 0 & -0.5 & 2 \end{pmatrix}
\]
Corollary: SOCP for \( c \)-optimality [Sag09,10]

If \( \mu \) is the optimal dual variable associated to the constraints of

\[
\max_{\mathbf{z} \in \mathbb{R}^m} \mathbf{c}^T \mathbf{z} \\
\forall i \in [s], \quad \|A_i \mathbf{z}\| \leq 1
\]

Then, \( \mathbf{w} := \frac{\mu^*}{\sum_{k=1}^{S} \mu_k^*} \) is a \( \mathbf{c} \)-optimal design.
Corollary: SOCP for $c$–optimality [Sag09,10]

If $\mu$ is the optimal dual variable associated to the constraints of

$$\begin{align*}
\max_{z \in \mathbb{R}^m} & \quad c^T z \\
\text{s.t.} & \quad A_i z \leq 1, \quad \forall i \in [s],
\end{align*}$$

Then, $w := \frac{\mu^*}{\sum_{k=1}^s \mu_k^*}$ is a $c$–optimal design.

SOCP for $A$–optimality [Sag10]

$$\begin{align*}
\max_{U \in \mathbb{R}^{m \times m}} & \quad \text{trace } U \\
\text{s.t.} & \quad A_i U \leq 1, \quad \forall i \in [s],
\end{align*}$$
Outline

1. Optimal Design of Experiments
2. Application to Network Monitoring
3. Algorithms for Optimal Experimental Design
   - Classic algorithms
   - Recent developments
   - New search directions
Thank you for your attention