Distributionally Robust Optimal Designs

Guillaume Sagnol

TU Berlin

Latest Advances in the Theory and Applications of Design and Analysis of Experiments,
BIRS Workshop, Banff, August 8, 2017

Outline

Design of Experiment

- $X \subset \mathbb{R}^n$: compact design space

An experiment with N trials is defined by a *design*

$$\xi = \left\{ \begin{array}{ccc} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_s \\ n_1 & \cdots & n_s \end{array} \right\},$$

where

- $\boldsymbol{x}_i \in X$ is the ith *support point* of the design
- $n_i \in \mathbb{N}$ is the replication at the ith design point
- $\sum_{i=1}^{s} n_i = N$.

Design of Experiment

- $X \subset \mathbb{R}^n$: compact design space

When $N \to \infty$, we can consider *approximate designs*:

$$\xi = \left\{ \begin{array}{ccc} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_s \\ w_1 & \cdots & w_s \end{array} \right\},$$

where

- $w_i \in \mathbb{R}_+$ is the proportion of the total number of trials at ith design point
- $\boldsymbol{x}_i \in X$ is a *support point* of the design iff $w_i > 0$
- $\sum_{i=1}^{s} w_i = 1$.

We denote by Ξ the set of all approximate designs

The Linear Model

We assume the following model:
A trial at the design point $\boldsymbol{x} \in X$ provides an observation

$$y = f(\boldsymbol{x})^T \boldsymbol{\theta} + \epsilon,$$

where

- $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ is an *unknown* vector of parameters;
- $f : X \mapsto \mathbb{R}^m$ is known;
- $\mathbb{E}[\epsilon] = \mathbf{0}, \quad \mathbb{V}[\epsilon] = \sigma^2$ (a known constant), and the noises ϵ, ϵ' of two distinct trials are uncorrelated.

Definition

The Fisher information matrix (FIM) of a design $\xi \in \Xi$ is

$$M(\xi) := \sum_{i=1}^{s} w_i \, f(\boldsymbol{x}_i) \, f(\boldsymbol{x}_i)^T \in \mathbb{S}_m^+.$$

Design for Estimation or Prediction

GOAL: Select a design $\xi \in \Xi$, such that

1. The vector $\boldsymbol{\theta}$ can be *estimated* with the best possible accuracy
2. OR, such that the function $\eta : \boldsymbol{x} \to f(\boldsymbol{x})^T\boldsymbol{\theta}$ can be *predicted* with the best possible accuracy

- These goals are essentially multicriterial (there are several θ_j's and many $\boldsymbol{x}$'s).
- So an appropriate scalarization is required.

Standard Optimality Criterions

- Designs for optimal estimation of θ:
  - $D-$**Optimality:** maxmize Determinant of information matrix $\leftrightarrow$ min. volume of conf. ellipsoids for θ.

Standard Optimality Criterions

- Designs for optimal estimation of θ:
  - $D-$**Optimality:** maxmize Determinant of information matrix
    $\leftrightarrow$ min. volume of conf. ellipsoids for θ.
  - $A-$**Optimality:** minimize trace of inverse of information matrix
    $\leftrightarrow$ min. diagonal of conf. ellipsoids for θ.

Standard Optimality Criterions

- Designs for optimal estimation of $\boldsymbol{\theta}$:
  - $D-$**Optimality:** maxmize Determinant of information matrix
    $\leftrightarrow$ min. volume of conf. ellipsoids for $\boldsymbol{\theta}$.
  - $A-$**Optimality:** minimize trace of inverse of information matrix
    $\leftrightarrow$ min. diagonal of conf. ellipsoids for $\boldsymbol{\theta}$.
  - A_K-**Optimality:** minimize $\mathrm{trace}\, K^T M(\xi)^{-1} K$
    $\leftrightarrow$ min. diagonal of conf. ellipsoids for the estimation of $K^T\boldsymbol{\theta}$.

Standard Optimality Criterions

- Designs for optimal estimation of $\boldsymbol{\theta}$:
  - $D-$**Optimality:** maxmize Determinant of information matrix
    $\leftrightarrow$ min. volume of conf. ellipsoids for $\boldsymbol{\theta}$.
  - $A-$**Optimality:** minimize trace of inverse of information matrix
    $\leftrightarrow$ min. diagonal of conf. ellipsoids for $\boldsymbol{\theta}$.
  - A_K-**Optimality:** minimize $\text{trace}\, K^T M(\xi)^{-1} K$
    $\leftrightarrow$ min. diagonal of conf. ellipsoids for the estimation of $K^T\boldsymbol{\theta}$.

- Designs for prediction of $y(\boldsymbol{x})$ at unsampled $\boldsymbol{x}$
  - $G-$**Optimality:** minimize worst-case prediction variance

$$\min_{\xi} \max_{\boldsymbol{x} \in X} \rho(\boldsymbol{x})$$

Standard Optimality Criterions

- Designs for optimal estimation of $\boldsymbol{\theta}$:
  - $D-$**Optimality:** maxmize Determinant of information matrix
    $\leftrightarrow$ min. volume of conf. ellipsoids for $\boldsymbol{\theta}$.
  - $A-$**Optimality:** minimize trace of inverse of information matrix
    $\leftrightarrow$ min. diagonal of conf. ellipsoids for $\boldsymbol{\theta}$.
  - A_K-**Optimality:** minimize $\mathrm{trace}\, K^T M(\xi)^{-1} K$
    $\leftrightarrow$ min. diagonal of conf. ellipsoids for the estimation of $K^T\boldsymbol{\theta}$.

- Designs for prediction of $y(\boldsymbol{x})$ at unsampled $\boldsymbol{x}$
  - $G-$**Optimality:** minimize worst-case prediction variance

$$\min_{\xi} \max_{\boldsymbol{x} \in X} \rho(\boldsymbol{x})$$

  - $I_\mu-$**Optimality:** minimize integrated prediction variance

$$\min_{\xi} \int_{\boldsymbol{x} \in X} \rho(\boldsymbol{x}) d\mu(\boldsymbol{x})$$

Equivalence Theorems

Theorem

A design is $G-$optimal iff it is $D-$optimal.

Theorem

Let $\boldsymbol{H} = \int_{\boldsymbol{x} \in X} f(\boldsymbol{x}) f(\boldsymbol{x})^T d\mu(\boldsymbol{x})$, and take any factorization $\boldsymbol{H} = \boldsymbol{K}\boldsymbol{K}^T$. Then, a design is $I_\mu-$optimal iff it is A_K-optimal.

Roughly speaking, all design problems (for prediction or estimation) reduce to the maximization of a function of the form $\Phi(M(\xi))$, where Φ is a concave design criterion.

The Nonlinear Model

Now, we assume that a trial at $\boldsymbol{x}$ yields a response

$$y = \eta(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon,$$

where $\eta : X \times \Theta \mapsto \mathbb{R}$ is a known function, and we define the sensitivity function

$$f(\boldsymbol{x}, \boldsymbol{\theta}) := \frac{\partial \eta}{\partial \boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}) \in \mathbb{R}^m$$

Local FIM

The FIM of a design $\xi \in \Xi$ now depends on $\boldsymbol{\theta} \in \Theta$:

$$M(\xi; \boldsymbol{\theta}) := \sum_{i=1}^{s} w_i\, f(\boldsymbol{x}_i; \boldsymbol{\theta})\, f(\boldsymbol{x}_i; \boldsymbol{\theta})^T \in \mathbb{S}_m^+$$

Remark: similar situation for the generalized linear model:

$$y \in \{0, 1\}, \qquad \mathbb{P}[y = 1] = \eta(\boldsymbol{x}, \boldsymbol{\theta}).$$

Dealing with parameter-dependency (1/2)

Given a design criterion $\Phi : \mathbb{S}_m^+ \mapsto \mathbb{R}$,

- Local optimal design at $\boldsymbol{\theta} \in \Theta$:

$$\max_{\xi \in \Xi} \Phi(M(\xi; \boldsymbol{\theta}))$$

- (Pseudo-)Bayesian optimal design:
  Given a prior π (probability measure over Θ),

$$\max_{\xi \in \Xi} \int_{\boldsymbol{\theta} \in \Theta} \Phi(M(\xi; \boldsymbol{\theta}))\, d\pi(\boldsymbol{\theta})$$

- Maximin Optimal Design

$$\max_{\xi \in \Xi} \min_{\boldsymbol{\theta} \in \Theta} \Phi(M(\xi; \boldsymbol{\theta}))$$

Dealing with parameter-dependency (2/2)

Standardized versions of these criterions have also been considered. Define the local efficiency of a design as

$$\mathrm{eff}(\xi;\boldsymbol{\theta}) := \frac{\Phi(M(\xi;\boldsymbol{\theta}))}{\sup_{\xi^* \in \Xi} \Phi(M(\xi^*;\boldsymbol{\theta}))} \in [0,1].$$

- Standardized Bayesian optimal design:
  Given a prior π (probability measure over Θ),

$$\max_{\xi \in \Xi} \int_{\boldsymbol{\theta} \in \Theta} \mathrm{eff}(\xi;\boldsymbol{\theta})\, d\pi(\boldsymbol{\theta})$$

- Standardized Maximin Optimal Design:

$$\max_{\xi \in \Xi} \min_{\boldsymbol{\theta} \in \Theta} \mathrm{eff}(\xi;\boldsymbol{\theta})$$

The Conic Programming approach

- When $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s\}$ is finite, the optimal design problem reduces to finding the vector of weights $\boldsymbol{w} \in \mathbb{R}^s$ of the design.
  $\rightarrow$ This is a convex optimization problem.
- A conic programming problem is a linear optimization problem over a convex cone $\mathcal{K}$
- Interior Point Methods are algorithms that are efficient both in theory and in practice, in particular for the following cones
  - $\mathcal{K} = \mathbb{R}_+^n$: Linear Programming (LP)
  - $\mathcal{K} = \mathcal{L}_n$: Second Order Cone Programming (SOCP)
  - $\mathcal{K} = \mathbb{S}_+^n$: Semidefinite Programming (SDP)

Conic-representability

- We say that a concave function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is $\mathcal{K}$-representable if its hypograph

$$\text{hypo } f := \{(\boldsymbol{x}, t) \in \mathbb{R}^{n+1} : f(\boldsymbol{x}) \geq t\}$$

  is equal to the projection of a set of the form $\{\boldsymbol{z} : A\boldsymbol{z} - \boldsymbol{b} \in \mathcal{K}\}$.
- The optimal design problem (for the linear model) can be reformulated as a conic optimization problem over $\mathcal{K}$ if the criterion Φ is $\mathcal{K}-$representable.
- Conic representability of design criterions:

criterion	E_K	A_K	D_K	$\boldsymbol{c}$	$\Phi_{p,K}$	$(p \leq 1, p \in \mathbb{Q})$
SDP	X	X	X	X	X	
SOCP	?	X	X	X	?	
LP				X		

Example: A-optimality

$$\Phi_A(M) := (\text{trace}\, M^{-1})^{-1}$$

Semidefinite representation of Φ_A:

$$\Phi_A(M) \geq t \iff \exists Y \in \mathbb{S}_m : \text{trace}\, Y \leq t \text{ and } \begin{bmatrix} Y & tI \\ tI & M \end{bmatrix} \succeq 0.$$

Example: A-optimality

$$\Phi_A(M) := (\operatorname{trace} M^{-1})^{-1}$$

Semidefinite representation of Φ_A:

$$\Phi_A(M) \geq t \iff \exists Y \in \mathbb{S}_m : \operatorname{trace} Y \leq t \text{ and } \begin{bmatrix} Y & tI \\ tI & M \end{bmatrix} \succeq 0.$$

A-optimality SDP:

$$\begin{aligned} \max_{\boldsymbol{w},t,Y} \quad & t \\ s.t. \quad & \operatorname{trace} Y \leq t \\ & \begin{bmatrix} Y & tI \\ tI & M(\boldsymbol{w}) \end{bmatrix} \succeq 0 \\ & \sum_i w_i = 1, \; \boldsymbol{w} \geq \mathbf{0}. \end{aligned}$$

Conic Programming Approach to DoE

- Maxdet and SDP formulations [e.g. Boyd & Vandenberghe, 2004]
- SDP-approach to compute criterion-robust designs [Harman, 2004]
- (MI)SOCP formulations for approximate (exact) $A-$ and $D-$optimality [S., 2011], [S. & Harman, 2015]
- SDP-approach to find support points in rational models [Papp, 2012]
- SDP formulation for Φ_p-optimality [S., 2013]
- Extented formulation for Bayesian Designs [Duarte, Wong, 2015]
- Extented formulation for Maximin Designs [Duarte, S., Wong, submitted]

Outline

Optimization under uncertainty

Terminology used in OR community

- $\boldsymbol{x}$: decision variable
- X decision space
- $\boldsymbol{\theta}$: uncertain parameter, with *nominal value* $\bar{\boldsymbol{\theta}}$.
- Θ : uncertainty set
- $F(\boldsymbol{x}, \theta)$: objective function (revenue)

- **Nominal (deterministic) Problem:**

$$\max_{\boldsymbol{x} \in X} \; F(\boldsymbol{x}, \bar{\boldsymbol{\theta}})$$

- **Stochastic Programming:**

$$\max_{\boldsymbol{x} \in X} \; \mathbb{E}_{\boldsymbol{\theta}} \, F(\boldsymbol{x}, \boldsymbol{\theta})$$

- **Robust Optimization:**

$$\max_{\boldsymbol{x} \in X} \; \min_{\boldsymbol{\theta} \in \Theta} \; F(\boldsymbol{x}, \boldsymbol{\theta})$$

Distributionally Robust Optimization

- Often, only a few samples from the uncertain parameter are available (e.g., historical data).
- This may not be enough to characterize exactly the distribution of $\boldsymbol{\theta}$.
- However, this data can be used to obtain (probabilistic) bounds on the expected value or variance of $\boldsymbol{\theta}$, or on the probability that $\boldsymbol{\theta} \in \Theta' \subset \Theta$.

Definition

Given a family $\mathcal{P}$ of probability distributions for the parameter $\boldsymbol{\theta}$, the *distributionally robust counterpart* (of the deterministic optimization problem) is

$$\max_{\boldsymbol{x} \in X} \min_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}} F(\boldsymbol{x}, \boldsymbol{\theta})$$

Review of main developments

- Introduced by Scarf (1958) for the Newsvendor Problem
- A lot of advances in the last decade, with the raise of Conic Programming (e.g. El Ghaoui et. al, 2003)
- When $F(x, \boldsymbol{\theta})$ is convex w.r.t. x and the ambiguity set $\mathcal{P}$ is defined through expected value of functions of $\boldsymbol{\theta}$, DRO reduces to a semi-infinite convex program
- Delage & Ye's seminal work (2010):
  - "Recipe" to construct an ambiguity set $\mathcal{P}$ from historical samples of $\boldsymbol{\theta}$, with theoretical foundations
  - If $\boldsymbol{\theta} \mapsto F(x, \boldsymbol{\theta})$ is concave and $x \mapsto F(x, \boldsymbol{\theta})$ is convex, separations oracles are provided, Θ is convex, then DRO is *tractable*.
  - If moreover $\boldsymbol{\theta} \mapsto F(x, \boldsymbol{\theta})$ is PWL and Θ is a polytope or an ellipsoid, the DRO problem reduces to an SDP.

Outline

Distributionally Robust Optimal Designs

Given a design criterion Φ and a family $\mathcal{P}$ of priors for the unknown vector of parameters $\boldsymbol{\theta}$, a design $\xi \in \Xi$ is called *distributionally robust optimal* (DRO) if it maximizes

$$\min_{\pi \in \mathcal{P}} \int_{\boldsymbol{\theta} \in \Theta} \Phi(M(\xi; \boldsymbol{\theta}))\, d\pi(\boldsymbol{\theta}).$$

Special cases:

- If $\mathcal{P} = \{\pi\}$ is a singleton:
  DRO design $\longleftrightarrow$ Bayesian optimal design
- If $\mathcal{P} = \{\mathbb{P} \text{ prob. measure} : \mathbb{P}(\Theta) = 1\}$:
  DRO design $\longleftrightarrow$ Maximin optimal design

A simple example

We first assume that Θ is finite:

$$\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}.$$

Consider the following family of priors:
Given $\boldsymbol{\theta} \in \Theta$ and $\Sigma \succ 0$,

$$\mathcal{P} = \left\{ \boldsymbol{p} \in \mathbb{R}^N_+ : \quad \begin{array}{l} \sum_k p_k = 1, \\ \sum_k p_k \boldsymbol{\theta}_k = \bar{\boldsymbol{\theta}}, \\ \sum_k p_k (\boldsymbol{\theta}_k - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_k - \bar{\boldsymbol{\theta}})^T = \Sigma \end{array} \right\}$$

SDP formulation: example

DRO-design:

$$\max_{\xi \in \Xi} \underbrace{\min_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}}[\Phi(M(\xi, \boldsymbol{\theta}))]}_{\Phi_{\mathrm{DRO}}(\xi)}$$

The inner optimization problem is a Linear Program (LP):

$$\begin{aligned} \Phi_{\mathrm{DRO}}(\xi) = \min_{\boldsymbol{p} \geq \boldsymbol{0}} \quad & \sum_k p_k \Phi(M(\xi; \boldsymbol{\theta_i})) \\ s.t. \quad & \sum_k p_k = 1, \\ & \sum_k p_k \boldsymbol{\theta}_k = \bar{\boldsymbol{\theta}}, \\ & \sum_k p_k \underbrace{(\boldsymbol{\theta}_k - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_k - \bar{\boldsymbol{\theta}})^T}_{V_k} = \Sigma \end{aligned}$$

SDP formulation: example

DRO-design:

$$\max_{\xi \in \Xi} \underbrace{\min_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}}[\Phi(M(\xi, \boldsymbol{\theta}))]}_{\Phi_{\mathrm{DRO}}(\xi)}$$

By the strong LP-duality theorem,

$$\begin{aligned} \Phi_{\mathrm{DRO}}(\xi) = \max_{\lambda \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^m, \Lambda \in \mathbb{S}^m} \quad & \lambda + \boldsymbol{\mu}^T \bar{\boldsymbol{\theta}} + \langle \Lambda, \Sigma \rangle \\ s.t. \quad & \lambda + \boldsymbol{\mu}^T \boldsymbol{\theta}_k + \langle \Lambda, V_k \rangle \leq \Phi(M(\xi; \boldsymbol{\theta}_k)) \\ & (\forall k = 1, \ldots, N) \end{aligned}$$

SDP formulation: example

DRO-design:

$$\max_{\xi \in \Xi} \underbrace{\min_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}}[\Phi(M(\xi, \boldsymbol{\theta}))]}_{\Phi_{\mathrm{DRO}}(\xi)}$$

By the strong LP-duality theorem,

$$\begin{aligned} \Phi_{\mathrm{DRO}}(\xi) = \max_{\lambda \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^m, \Lambda \in \mathbb{S}^m} \quad & \lambda + \boldsymbol{\mu}^T \bar{\boldsymbol{\theta}} + \langle \Lambda, \Sigma \rangle \\ s.t. \quad & \lambda + \boldsymbol{\mu}^T \boldsymbol{\theta}_k + \langle \Lambda, V_k \rangle \leq \Phi(M(\xi; \boldsymbol{\theta}_k)) \\ & (\forall k = 1, \ldots, N) \end{aligned}$$

Finally, maximizing the above expression with respect to $\xi \in \Xi$ is an SDP when X is finite and Φ is SDP-representable.

Simple Example: the general case

$$\mathcal{P} = \left\{ \pi \text{ prob. measure} : \begin{array}{l} \int_\Theta d\pi(\boldsymbol{\theta}) = 1, \\ \int_\Theta \boldsymbol{\theta}\, d\pi(\boldsymbol{\theta}) = \bar{\boldsymbol{\theta}}, \\ \int_\Theta (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T\, d\pi(\boldsymbol{\theta}) = \Sigma \end{array} \right\}$$

Theorem

A design $\xi \in \Xi$ is DRO iff there exists a dual probability measure $\pi \in \mathcal{P}$, as well as $(\lambda, \boldsymbol{\mu}, \Lambda) \in \mathbb{R} \times \mathbb{R}^m \times \mathbb{S}^m$ such that

- ξ is Bayesian optimal for π
- $\forall \boldsymbol{\theta} \in \Theta, \quad \lambda + \boldsymbol{\mu}^T \boldsymbol{\theta} + (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T \Lambda (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \leq \Phi(M(\xi; \boldsymbol{\theta}))$

Moreover, the above inequality becomes an equality at the support points of π.

Convex tractable sets of distribution families

The framework we propose is working for families $\mathcal{P}$ of probability distributions $\mathbb{P}$ satisfying $\mathbb{P}(\Theta) = 1$, as well as constraints of the form

- $\mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}}[\psi_i(\boldsymbol{\theta})] = 0$,
  where $\psi_i : \Theta \mapsto \mathbb{R}$ is a continuous function

Convex tractable sets of distribution families

The framework we propose is working for families $\mathcal{P}$ of probability distributions $\mathbb{P}$ satisfying $\mathbb{P}(\Theta) = 1$, as well as constraints of the form

- $\mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}}[\psi_i(\boldsymbol{\theta})] = 0$,
  where $\psi_i : \Theta \mapsto \mathbb{R}$ is a continuous function
- $\mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}}[\Psi_j(\boldsymbol{\theta})] \succeq 0$,
  where $\Psi_j : \Theta \mapsto \mathbb{S}_{n_j}$ is a continuous function

(and we assume a Slater-type condition holds).

Convex tractable sets of distribution families

The framework we propose is working for families $\mathcal{P}$ of probability distributions $\mathbb{P}$ satisfying $\mathbb{P}(\Theta) = 1$, as well as constraints of the form

- $\mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}}[\psi_i(\boldsymbol{\theta})] = 0$,
  where $\psi_i : \Theta \mapsto \mathbb{R}$ is a continuous function
- $\mathbb{E}_{\boldsymbol{\theta} \sim \mathbb{P}}[\Psi_j(\boldsymbol{\theta})] \succeq 0$,
  where $\Psi_j : \Theta \mapsto \mathbb{S}_{n_j}$ is a continuous function

(and we assume a Slater-type condition holds).

In particular, this allows constraints of the form

- $\mathbb{E}[\boldsymbol{\theta}]$ belongs to a convex set

Convex tractable sets of distribution families

The framework we propose is working for families $\mathcal{P}$ of probability distributions $\mathbb{P}$ satisfying $\mathbb{P}(\Theta) = 1$, as well as constraints of the form

- $\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[\psi_i(\boldsymbol{\theta})] = 0$,
  where $\psi_i : \Theta \mapsto \mathbb{R}$ is a continuous function
- $\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[\Psi_j(\boldsymbol{\theta})] \succeq 0$,
  where $\Psi_j : \Theta \mapsto \mathbb{S}_{n_j}$ is a continuous function

(and we assume a Slater-type condition holds).

In particular, this allows constraints of the form

- $\mathbb{E}[\boldsymbol{\theta}]$ belongs to a convex set
- Bounds on the probability that $\boldsymbol{\theta} \in \Theta'$, where $\Theta' \subseteq \Theta$

Convex tractable sets of distribution families

The framework we propose is working for families $\mathcal{P}$ of probability distributions $\mathbb{P}$ satisfying $\mathbb{P}(\Theta) = 1$, as well as constraints of the form

- $\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[\psi_i(\boldsymbol{\theta})] = 0$,
  where $\psi_i : \Theta \mapsto \mathbb{R}$ is a continuous function
- $\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[\Psi_j(\boldsymbol{\theta})] \succeq 0$,
  where $\Psi_j : \Theta \mapsto \mathbb{S}_{n_j}$ is a continuous function

(and we assume a Slater-type condition holds).

In particular, this allows constraints of the form

- $\mathbb{E}[\boldsymbol{\theta}]$ belongs to a convex set
- Bounds on the probability that $\boldsymbol{\theta} \in \Theta'$, where $\Theta' \subseteq \Theta$
- $\mathbb{V}[\boldsymbol{\theta}] \succeq \Sigma_0$ (w.r.t. Loewner ordering)
  Indeed, this is equivalent to $\mathbb{E}\begin{bmatrix} (\boldsymbol{\theta}\boldsymbol{\theta}^T - \Sigma_0) & \boldsymbol{\theta} \\ \boldsymbol{\theta}^T & 1 \end{bmatrix} \succeq 0$

Semi-infinite formulation for finite X

Let $\mathcal{P} =$

$$\left\{ \mathbb{P} \text{ prob. measure} : \begin{array}{ll} \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[1] = 1 & \\ \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[\psi_i(\boldsymbol{\theta})] = 0 & (i = 1, \ldots, p) \\ \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[\Psi_j(\boldsymbol{\theta})] \succeq 0 & (j = 1, \ldots, q) \end{array} \right\},$$

and assume that $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s\}$ is finite.

Then, the weights w_k of a DRO-design $\xi^* = \{\boldsymbol{x}_k, w_k\}$ solve the following semi-infinite SDP:

$$\begin{aligned} \max_{\boldsymbol{w},\lambda,\boldsymbol{\mu},\Lambda_j} \quad & \lambda \\ s.t. \quad & \Phi(M(\boldsymbol{w}, \boldsymbol{\theta})) \geq \lambda + \sum_i \mu_i \psi_i(\boldsymbol{\theta}) + \sum_j \langle \Lambda_j, \Psi_j(\boldsymbol{\theta}) \rangle, \\ & \qquad (\forall \boldsymbol{\theta} \in \Theta) \\ & \sum_k w_k = 1, \;\; \boldsymbol{w} \geq \mathbf{0} \\ & \Lambda_j \succeq 0 \; (j = 1, \ldots, q). \end{aligned}$$

Optimality conditions

$$\mathcal{P} = \left\{ \mathbb{P} \text{ prob. measure : } \begin{array}{ll} \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[1] = 1 & \\ \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[\psi_i(\boldsymbol{\theta})] = 0 & (i = 1, \ldots, p) \\ \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[\Psi_j(\boldsymbol{\theta})] \succeq 0 & (j = 1, \ldots, q) \end{array} \right\}.$$

Theorem

If the *ambiguity set* $\mathcal{P}$ contains a Slater-type point, then a design $\xi \in \Xi$ is DRO iff there exists a dual probability measure $\pi \in \mathcal{P}$, as well as

$$(\lambda, \boldsymbol{\mu}, \Lambda_1, \ldots, \Lambda_q) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{S}^{n_1} \times \cdots \times \mathbb{S}^{n_q}$$

such that

- ξ is Bayesian optimal for π
- $\forall \boldsymbol{\theta} \in \Theta, \quad \lambda + \sum_i \mu_i \psi_i(\boldsymbol{\theta}) + \sum_j \langle \Lambda_j, \Psi_j(\boldsymbol{\theta}) \rangle \leq \Phi(M(\xi; \boldsymbol{\theta}))$

Moreover, the above inequality becomes an equality at the support points of π.

Example: Delage & Ye's Ambiguity set

Ambiguity set of Delage and Ye for Data-Driven DRO:
Given some *estimates* $\boldsymbol{\mu}$ and Σ for the mean and variance of $\boldsymbol{\theta}$, and some confidence parameters $\gamma_1, \gamma_2 \geq 0$,

$$\mathcal{P} = \left\{ \mathbb{P} \text{ prob. measure} : \begin{array}{l} \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[1] = 1 \\ \mathbb{E}_{(\boldsymbol{\theta}\sim\mathbb{P}}[(\boldsymbol{\theta}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})] \leq \gamma_1 \\ \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})^T] \preceq (1+\gamma_2)\Sigma \end{array} \right\}.$$

Example: Delage & Ye's Ambiguity set

Ambiguity set of Delage and Ye for Data-Driven DRO:
Given some *estimates* $\boldsymbol{\mu}$ and Σ for the mean and variance of $\boldsymbol{\theta}$, and some confidence parameters $\gamma_1, \gamma_2 \geq 0$,

$$\mathcal{P} = \left\{ \mathbb{P} \text{ prob. measure} : \begin{array}{l} \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[1] = 1 \\ \mathbb{E}_{(\boldsymbol{\theta}\sim\mathbb{P}}[(\boldsymbol{\theta}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})] \leq \gamma_1 \\ \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})^T] \preceq (1+\gamma_2)\Sigma \end{array} \right\}.$$

Semi-infinite SDP:

$$\begin{aligned} \max_{\boldsymbol{w},\beta,Q} \quad & \lambda - \beta\gamma_1 - (1+\gamma_2)\langle Q, \Sigma\rangle \\ s.t. \quad & \Phi(M(\boldsymbol{w},\boldsymbol{\theta})) \geq \lambda - (\boldsymbol{\theta}-\boldsymbol{\mu})^T(\Sigma^{-1}+Q)(\boldsymbol{\theta}-\boldsymbol{\mu}) \qquad (\forall \boldsymbol{\theta} \in \Theta) \\ & \sum_k w_k = 1,\ \boldsymbol{w} \geq \mathbf{0} \\ & \beta \geq 0, Q \succeq 0. \end{aligned}$$

Example: Delage & Ye's Ambiguity set

Ambiguity set of Delage and Ye for Data-Driven DRO:
Given some *estimates* $\boldsymbol{\mu}$ and Σ for the mean and variance of $\boldsymbol{\theta}$, and some confidence parameters $\gamma_1, \gamma_2 \geq 0$,

$$\mathcal{P} = \left\{ \mathbb{P} \text{ prob. measure}: \begin{array}{l} \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[1] = 1 \\ \mathbb{E}_{(\boldsymbol{\theta}\sim\mathbb{P}}[(\boldsymbol{\theta}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})] \leq \gamma_1 \\ \mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{P}}[(\boldsymbol{\theta}-\boldsymbol{\mu})(\boldsymbol{\theta}-\boldsymbol{\mu})^T] \preceq (1+\gamma_2)\Sigma \end{array} \right\}.$$

SDP for A-optimality over a finite $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$:

$$\begin{aligned} \max_{\boldsymbol{w},\beta,Q,\boldsymbol{t},Y_k} \quad & \lambda - \beta\gamma_1 - (1+\gamma_2)\langle Q, \Sigma\rangle \\ s.t. \quad & t_k = \lambda - (\boldsymbol{\theta}_k - \boldsymbol{\mu})^T(\Sigma^{-1} + Q)(\boldsymbol{\theta}_k - \boldsymbol{\mu}) \geq \operatorname{tr} Y_k \\ & \begin{bmatrix} Y_k & t_k I \\ t_k I & M(\boldsymbol{w}, \boldsymbol{\theta}_k) \end{bmatrix} \succeq 0 \qquad (k = 1, \ldots, N) \\ & \sum_k w_k = 1,\ \boldsymbol{w} \geq \mathbf{0}, \beta \geq 0, Q \succeq 0. \end{aligned}$$

Asymptotic result

Let ξ^* be a DRO-design over a finite set of candidate points X.

Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ be an i.i.d. sample over Θ (from any *continuous* distribution), and denote by ξ_N the design computed by the SDP over $\Theta_N = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$.

Asymptotic result

Let ξ^* be a DRO-design over a finite set of candidate points X.

Let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ be an i.i.d. sample over Θ (from any *continuous* distribution), and denote by ξ_N the design computed by the SDP over $\Theta_N = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$.

Theorem [Xiu, Liu & Sun, 2017]

Under some regularity conditions, for any $\epsilon > 0$, there exists constants $C > 0$ and $\beta > 0$ such that for N sufficiently large,

$$\text{Prob}(|\Phi_{\text{DRO}}(\xi^*) - \Phi_{\text{DRO}}(\xi_N)| > \epsilon) \leq Ce^{-\beta N}.$$

A primal-dual cutting-plane algorithm

- Start with a discretization $\hat{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s\} \subset X$, and $\hat{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\} \subset \Theta$
- Repeat until convergence:
  1. Solve the finite-size SDP over $\hat{X}$ and $\hat{\Theta}$
  2. The SDP solver returns a design ξ^*, Lagrange multipliers $\lambda, (\mu_i)_{1 \leq i \leq p}, (\Lambda_j)_{1 \leq j \leq q}$, and the optimal dual variables yield a dual measure π^* supported by $\hat{\Theta}$.
  3. Find some points $\boldsymbol{\theta}$ violating

     $$\lambda + \sum_i \mu_i \psi_i(\boldsymbol{\theta}) + \sum_j \langle \Lambda_j, \Psi_j(\boldsymbol{\theta}) \rangle \leq \Phi(M(\xi^*; \boldsymbol{\theta}))$$

     and add them to $\hat{\Theta}$
  4. Find some points $\boldsymbol{x}$ violating

     $$\int_{\boldsymbol{\theta} \in \Theta} \big(D\Phi(\xi^*; \boldsymbol{\theta})[\boldsymbol{x}] - \Phi(\xi^*; \boldsymbol{\theta}) \big) d\pi^*(\boldsymbol{\theta}) \leq 0$$

     and add then to $\hat{X}$

Outline

Logistic Regression in Two Variables

Model:

- $X = [1, 21] \times [1, 21]$
- $\Theta = \{0.1\} \times [0, 0.3] \times [0, 0.4]$
- GLM with logit-link function:

$$\mathrm{Prob}[y(\boldsymbol{x}) = 1] = p(\boldsymbol{x}, \boldsymbol{\theta}) \frac{\exp(\theta_0 + x_1\theta_1 + x_2\theta_2)}{1 + \exp(\theta_0 + x_1\theta_1 + x_2\theta_2)}$$

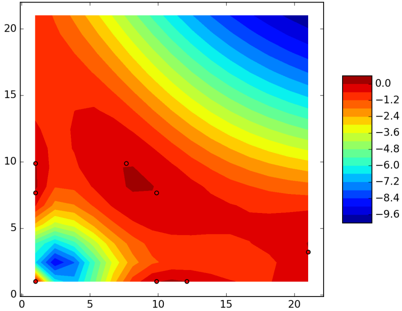
- $M(\delta_{\boldsymbol{x}}, \boldsymbol{\theta}) = p(\boldsymbol{x}, \boldsymbol{\theta})(1 - p(\boldsymbol{x}, \boldsymbol{\theta})) \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}^T$

We compute a Φ_A-DRO design over the family of priors s.t.:

- $\mathbb{E}[\boldsymbol{\theta}] = [0.1, 0.15, 0.2]$,
- $\mathbb{V}[\boldsymbol{\theta}] = \mathrm{diag}([0, 0.01, 0.01])$

Functions from the “equivalence theorem”

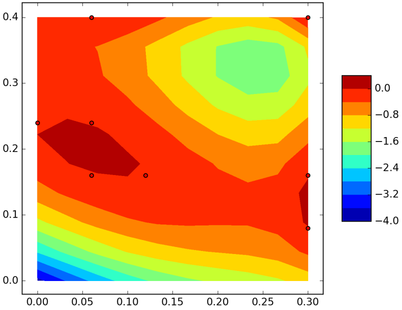
1st iteration



Plot of the function

$$\int_{\boldsymbol{\theta}\in\Theta} \big(D\Phi(\xi^*;\boldsymbol{\theta})[\boldsymbol{x}] - \Phi(\xi^*;\boldsymbol{\theta})\big)\, d\pi^*(\boldsymbol{\theta})$$

over $\boldsymbol{x} \in X$



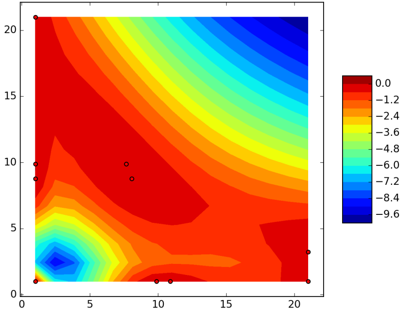
Plot of the function

$$\lambda + \boldsymbol{\mu}^T\boldsymbol{\theta} + (\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})^T\Lambda(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}) - \Phi(M(\xi^*;\boldsymbol{\theta}))$$

over $\boldsymbol{\theta} \in \Theta$

Functions from the “equivalence theorem”

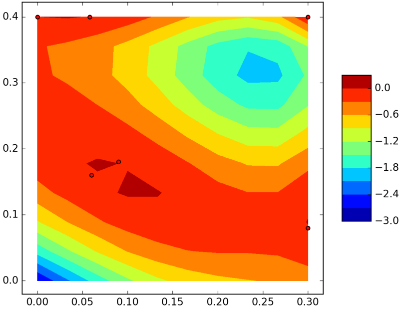
2nd iteration



Plot of the function

$$\int_{\theta \in \Theta} \left(D\Phi(\xi^*; \boldsymbol{\theta})[\boldsymbol{x}] - \Phi(\xi^*; \boldsymbol{\theta}) \right) d\pi^*(\boldsymbol{\theta})$$

over $\boldsymbol{x} \in X$



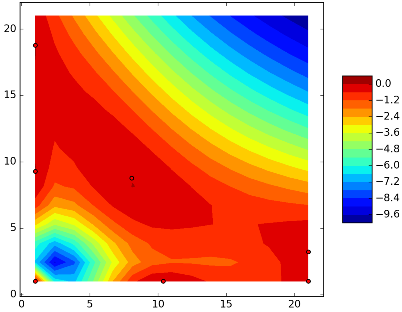
Plot of the function

$$\lambda + \boldsymbol{\mu}^T \boldsymbol{\theta} + (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T \Lambda (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) - \Phi(M(\xi^*; \boldsymbol{\theta}))$$

over $\boldsymbol{\theta} \in \Theta$

Functions from the “equivalence theorem”

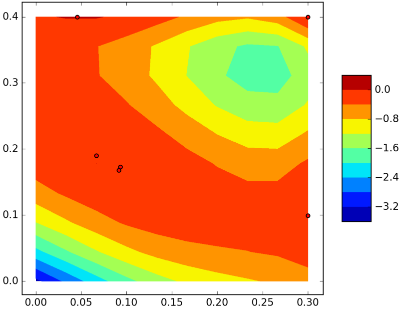
3rd iteration



Plot of the function

$$\int_{\boldsymbol{\theta}\in\Theta} \left(D\Phi(\xi^*;\boldsymbol{\theta})[\boldsymbol{x}] - \Phi(\xi^*;\boldsymbol{\theta})\right) d\pi^*(\boldsymbol{\theta})$$

over $\boldsymbol{x} \in X$

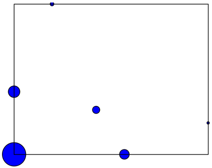


Plot of the function

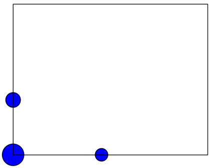
$$\lambda + \boldsymbol{\mu}^T\boldsymbol{\theta} + (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T \Lambda (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) - \Phi(M(\xi^*;\boldsymbol{\theta}))$$

over $\boldsymbol{\theta} \in \Theta$

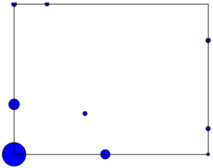
Optimal Designs



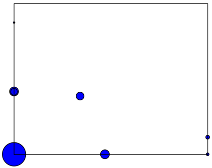
Bayes (uniform)



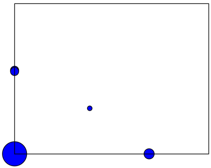
Maximin



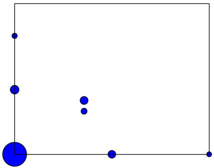
std. maximin



DRO ($\mathbb{V}[\theta_i]] = 0.01$)



DRO ($\mathbb{V}[\theta_i]] = 0.002$)



std.DRO ($\mathbb{V}[\theta_i]] = 0.01$)

Conclusion

- A new unifying framework to handle dependency to unknown parameters
- Flexibility to define the “ambiguity set”, partially overcomes drawbacks of Bayesian and Maximin approaches
- SDP formulation when X and Θ are discretized
- Primal-dual cutting-plane approach to find DRO-optimal design
- Approach can be extended to standardized design criterions

References

A few references (on DRO only):

- Scarf, H. 1958. A min-max solution of an inventory problem. K. Arrow, ed. Studies in the Mathematical Theory of Inventory and Production. Stanford University Press, Stanford, CA, 201–209.
- El Ghaoui, L., M. Oks, F. Oustry. 2003. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. Oper. Res. 51(4) 543–556.
- Delage, E. and Ye, Y., 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations research, 58(3), pp.595-612.
- Xu, H., Liu, Y., Sun, H. 2017. Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane methods. Math. Program., Ser. A, 2017. To Appear.