Mathematical Programming II	**Lecture 27**
ORIE 6310 Spring 2014	**May 6, 2014**
Scribe: Stephen Pallone	

Optimal Experimental Design Problem

$$\begin{aligned} \max \quad & \log\det XUX^T \\ \text{st.} \quad & e^T u = 1 \\ & u \geq 0, \end{aligned}$$

where $X = (x_1, \ldots, x_p)$ and $U = \text{Diag}(u)$. We aim to take the Lagrangian Dual, but first we reformulate this as

$$\begin{aligned} \max \quad & \log\det A \\ \text{st.} \quad & A - XUX^T = 0 \\ & e^T u = 1 \\ & u \geq 0 \\ & A \in \mathbb{S}^{m\times m}, \end{aligned} \tag{D}$$

where $\mathbb{S}^{m\times m}$ represents the space of symmetric $m \times m$ matrices. Similarly, $\mathbb{S}_+^{m\times m}$ denotes the positive semi-definite cone, and $\mathbb{S}_{++}^{m\times m}$ defines positive definite $m \times m$ matrices.

Trace Inner Product

Suppose $A, B \in \mathbb{R}^{m\times n}$. We can define an inner product as

$$\langle A, B\rangle = A \bullet B = \text{trace}\,(A^T B) = \sum_{i,j} a_{ij} b_{ij}.$$

Because $\text{trace}\,(UV) = \text{trace}\,(VU)$, the commutative property holds. Consider, for $A \in \mathbb{R}^{m\times m}$ with $\det A > 0$,

$$\begin{aligned} \phi(A) &= \log\det A, \\ \theta(\lambda) &:= \phi(A + \lambda e_i e_j^T) \\ &= \log\det(A + \lambda e_i e_j^T) \\ &= \log\left[\det(A)\left(1 + \lambda e_j^T A^{-1} e_i\right)\right] \\ &= \log\det A + \log\left(1 + \lambda e_j^T A^{-1} e_i\right). \end{aligned}$$

This tells us the directional derivative of ϕ is given by

$$\begin{aligned}\phi'(A; e_i e_j^T) &= e_j^T A^{-1} e_i \\ &= \mathrm{trace}\,(e_j^T A^{-1} e_i) \\ &= \mathrm{trace}\,(A^{-1} e_i e_j^T) \\ &= A^{-T} \bullet e_i e_j^T,\end{aligned}$$

and since $\{e_i e_j^T\}$ is a basis for $\mathbf{R}^{m\times m}$ and ϕ is smooth, we have

$$\phi'(A; D) = A^{-T} \bullet D.$$

Thus, $\nabla\phi(A) = A^{-T}$, and for A symmetric, $\nabla\phi(A) = A^{-1}$.

Aside: If we write $\psi(A) := \phi(A; D) = A^{-1} \bullet D$, where A, D are symmetric, we find $\psi'(A; E) = -\mathrm{trace}\,(A^{-1} D A^{-1} E)$. This is analogous to the second derivative, and can be used to show $-\log\det$ is convex in the positive semi-definite cone.

Lagrangian Dual

Now consider (D). This is equivalent to

$$\max_{A\in\mathbb{S}^{m\times m},\, u\geq 0}\ \min_{H\in\mathbb{S}^{m\times m},\, \lambda\in\mathrm{R}} \left(\log\det A - H \bullet A + H \bullet XUX^T - \lambda e^T u + \lambda\right).$$

Using the fact that $XUX^T = \sum u_i x_i x_i^T$, the dual is given by

$$\min_{H\in\mathbb{S}^{m\times m},\, \lambda\in\mathrm{R}}\ \max_{A\in\mathbb{S}^{m\times m},\, u\geq 0} \left([\log\det A - H \bullet A] + \sum_{i=1}^{p} u_i[x_i^T H x_i - \lambda] + \lambda\right).$$

Since the concave function $\log\det A - H \bullet A$ for positive semi-definite A is maximized by choosing $A = H^{-1}$ if H is positive definite. (If H is not positive definite, e.g., $v^T H v \leq 0$ for some $v \neq 0$, then choosing $A = I + \lambda vv^T$, with $\lambda \to \infty$, gives an infinite maximum.) Thus, we have

$$\min_{x_i^T H x_i \leq \lambda,\ H\in\mathbb{S}_{++}^{m\times m},\ \lambda\geq 0} \left([\log\det H^{-1} - H \bullet H^{-1}] + \lambda\right),$$

or if we simplify,

$$\begin{aligned}\min \quad & -\log\det H + \lambda - m \\ \text{st.} \quad & x_i^T H x_i \leq \lambda \quad \forall i = 1, \ldots, p \\ & \lambda \geq 0 \\ & H \in \mathbb{S}_{++}^{m\times m}.\end{aligned}$$

We can set $M = \frac{m}{\lambda}H$ to get

$$\begin{array}{rl} \min & -\log\det M - m\log\lambda + m\log m + \lambda - m \\ \text{st.} & x_i^T M x_i \le m \quad \forall i = 1,\ldots,p \\ & \lambda > 0, \quad H \in \mathbb{S}_{++}^{m\times m}. \end{array}$$

This separates the variables λ and M, and minimization over λ gives $\lambda = m$. This gives the final formulation of the dual problem

$$\begin{array}{rl} \min & -\log\det M \\ & x_i^T M x_i \le m \quad \forall i = 1,\ldots,p \\ & M \in \mathbb{S}_{++}^{m\times m}. \end{array} \tag{P}$$

The dual problem is analogous to finding the minimum volume central ellipsoid containing the points $\{x_i\}$.

D-optimality vs. G-optimality

We chose to minimize the determinant of the covariance matrix for $\hat{\beta}$, and this led to *D-optimality*. Also, if we made another test at the design point x_i, the variance of our estimate would be $x_i^T(XUX^T)^{-1}x_i$. So we might want to minimize $\max_i x_i^T(XUX^T)^{-1}x_i$ over $\{u \in \mathbf{R}^p : e^T u = 1,\ u \ge 0\}$. This criterion leads to *G-optimality*.

Proposition 1 *This is minimized by the same u that solves (D).*

Proof:

(a) For any feasible u,

$$\begin{aligned} \max_i x_i^T(XUX^T)^{-1}x_i &\ge \sum u_i x_i^T(XUX^T)^{-1}x_i \\ &= \sum \operatorname{trace}\left((XUX^T)^{-1}u_i x_i x_i^T\right) \\ &= \operatorname{trace}\left((XUX^T)^{-1}(XUX^T)\right) \\ &= m. \end{aligned}$$

(b) (Sketch) We can *achieve* this bound of m by choosing the optimal u from (D) so that $M = (XUX^T)^{-1}$ gives the maximum that equals m by duality. $\square$

Algorithms

We choose to solve (D), which also gives us an optimal solution to (P). Since $\log\det XUX^T$ is infinitely differentiable with nice expressions for its derivatives, we are tempted to use second-order methods for its solution. But every iteration is very expensive.

Consider instead coordinate ascent! Either increase $u^{(i)}$ or decrease $u^{(j)}$ at each iteration, then rescale. If we increase $u^{(i)}$ by λ, XUX^T increases by $\lambda x_i x_i^T$, a rank-one perturbation! So we can easily update $g(u) = \log\det XUX^T$ and $\nabla g(u) = (x_i^T(XUX^T)^{-1}x_i)_{i=1}^p$. Coordinate ascent with the correct choice of i or j is steepest ascent with respect to the L_1-norm.

This algorithm, with only increases in components, is due to Federov-Wynn (statistics) and Frank-Wolfe (optimization). The algorithm with increases *and* decreases is due to Atwood (statistics) and Wolfe (optimization). Khachiyan also contributed with a complexity analysis of the algorithm with just increases (remember the smallest ellipsoid problem in (P)). Ahipasaoglu-Sun-Todd proved linear convergence of the algorithm with both increases and decreases.

Final Remarks on the Course

Linear Complementarity Problem

- Pivoting algorithm "like" simplex method, but with no guiding objective function.
- Purely combinatorial proof of finite convergence; for suitable problems, we get either complementary solution or certificate of infeasibility.

Complexity of Pivoting Algorithms

- Neighborly polytopes, bound on diameters of polyhedra.
- Polynomial expected behavior of certain pivoting algorithms.
- Smoothed complexity.

Informational Complexity of Non-Linear Optimization Problems

Impossible to efficiently approximate the minimum of *non-convex* functions, or approximate the *minimizer* of convex functions. But we can approximate the *minimum* of convex functions.

- **Low-dimension, High Accuracy:** Method of Centers of Gravity, Ellipsoid Algorithm, Method of Inscribed Ellipsoids.
- **High-dimension, Low Accuracy:** (Sub)-gradient methods.

Interpretable Duals

- Regression;
- Data Classification;
- Optimal Experimental Design.

The End. $\square$