Mathematical Programming II **Lecture 26**
ORIE 6310 Spring 2014 **May 1, 2014**
Scribe: Daniel Muñoz

More on Data Classification

Last lecture, we focused on a data classification problem based on the L_∞-norm of $1/r_i$:

$$(\mathcal{P}_\infty) \qquad \begin{aligned} \min\ & Ce^\top \xi + \max \frac{1}{r_i} \\ & YX^\top w + \beta y + \xi - r = 0 \\ & \|w\|_2 \le 1,\ \xi \ge 0,\ r \ge 0 \end{aligned} \qquad\qquad (\mathcal{D}_\infty) \qquad \begin{aligned} \max\ & -\|XY\alpha\|_2 + 2\sqrt{e^\top \alpha} \\ & y^\top \alpha = 0 \\ & 0 \le \alpha \le Ce. \end{aligned}$$

Instead, today we'll use an L_1-norm on $1/r_i$. That is,

$$\begin{aligned} \min\ & Ce^\top \xi + \sum_{i=1}^n \frac{1}{r_i} \\ & YX^\top w + \beta y + \xi - r = 0 \\ & \|w\|_2 \le 1,\ \xi \ge 0,\ r \ge 0. \end{aligned}$$

As before, we use ρ, σ, but now in $\mathbf{R}^n$, and set $r_i = \rho_i - \sigma_i$, $1/r_i = \rho_i + \sigma_i$ and impose $\rho_i \ge \left\| \begin{pmatrix} \sigma_i \\ 1 \end{pmatrix} \right\|$ for each i.

This gives the conic problem $(\mathcal{P})$

$$\begin{aligned} \min_{\omega, w, \beta, \xi, \rho, \sigma, \tau}\ & Ce^\top \xi + e^\top \rho + e^\top \sigma \\ & YX^\top w + \beta y + \xi - \rho + \sigma = 0 \\ & \omega = 1 \\ & \tau = e \\ & \begin{pmatrix} \omega \\ w \end{pmatrix} \in K_2^{1+d}, \beta \in \mathbf{R}, \xi \ge 0, \begin{pmatrix} \rho_i \\ \sigma_i \\ \tau_i \end{pmatrix} \in K_2^{1+2}, i = 1, \ldots, n. \end{aligned}$$

Taking the conic dual and simplifying, we get $(\mathcal{D})$

$$\begin{aligned} \max & -\|XY\alpha\|_2 + 2e^\top \sqrt{\alpha} \\ & y^\top \alpha = 0 \\ & 0 \le \alpha \le Ce, \end{aligned}$$

where $\sqrt{\alpha} = (\sqrt{\alpha_i})$. Note that the optimal α will have all strictly positive entries.

Now, assume as in the last lecture that the data is separable, that C is sufficiently large so that $\xi = 0$ in $(\mathcal{P})$, and that $\alpha \leq Ce$ is non-binding in $(\mathcal{D})$. We write $\alpha_+ = \gamma\overline{\alpha}_+$, $\alpha_- = \gamma\overline{\alpha}_-$ with $e_+^\top\overline{\alpha}_+ = e_-^\top\overline{\alpha}_- = 1$, to get

$$\max_{\overline{\alpha}_+,\overline{\alpha}_-,\gamma} -\gamma \left\|X_+\overline{\alpha}_+ - X_-\overline{\alpha}_-\right\|_2 + 2\sqrt{\gamma}(e_+^\top\sqrt{\overline{\alpha}_+} + e_-^\top\sqrt{\overline{\alpha}_-}),$$

and maximizing γ first, we get

$$\gamma = \frac{(e_+^\top\sqrt{\overline{\alpha}_+} + e_-^\top\sqrt{\overline{\alpha}_-})^2}{\left\|X_+\overline{\alpha}_+ - X_-\overline{\alpha}_-\right\|_2^2},$$

and the dual becomes

$$\max_{\overline{\alpha}_+,\overline{\alpha}_-} \frac{(\overline{e}_+^\top\sqrt{\overline{\alpha}_+} + \overline{e}_-^\top\sqrt{\overline{\alpha}_-})^2}{\left\|X_+\overline{\alpha}_+ - X_-\overline{\alpha}_-\right\|_2} \quad \equiv \quad \min_{\overline{\alpha}_+,\overline{\alpha}_-} \frac{\left\|X_+\overline{\alpha}_+ - X_-\overline{\alpha}_-\right\|_2}{(\overline{e}_+^\top\sqrt{\overline{\alpha}_+} + \overline{e}_-^\top\sqrt{\overline{\alpha}_-})^2}$$

(contrast this with $\min_{\overline{\alpha}_+,\overline{\alpha}_-} \frac{\|X_+\overline{\alpha}_+ - X_-\overline{\alpha}_-\|_2}{2}$ last time). So now we minimize the distance between convex combinations of positive and negative points, weighted by the square of the sum of the square roots of the weights.

Again, as long as $XY\alpha \neq 0$ (which holds for separable data), the primal variable w will be $XY\alpha/\|XY\alpha\|_2$.

Optimal Experimental Design (Kiefer & Wolfowitz at Cornell, and others)
Consider the linear regression model

$$Y = x(t)^\top\beta + \epsilon,$$

where Y is the dependent variable, $t \in T$ is/are the independent variables, x is a known function from T to $\mathbf{R}^m$, and ϵ is an error vector normally distributed $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Here, the choice of a linear dependence on $x(t)$ is not as restrictive as it seems since nonlinear dependencies can be modeled, e.g., by defining $x(t) = (1, t, t^2, t^3)^\top$.

Now, suppose we observe Y at values $t_1, \ldots, t_n$ and get

$$y = (y_1, \ldots, y_n)^\top \approx X^\top\beta \ + (\epsilon_1, \ldots, \epsilon_n)^\top =: X^\top\beta + \overrightarrow{\epsilon},$$

where $X = [x_1, \ldots, x_n] = [x(t_1), \ldots, x(t_n)]$. A natural estimate for β is the least-squares estimator: $\hat{\beta} = (XX^\top)^{-1}Xy$, where y is a sample from $X^\top\beta + \overrightarrow{\epsilon}$, and so $\hat{\beta}$ is a sample of $(XX^\top)^{-1}X(X^\top\beta + \overrightarrow{\epsilon})$ or $\beta + (XX^\top)^{-1}X\overrightarrow{\epsilon}$. Here, we assume X has rank m.

Thus, $\mathbb{E}(\hat{\beta}) = \beta$ which means the estimate is unbiased, and its variance is the matrix

$$\begin{aligned}\mathbb{E}((\hat{\beta}-\beta)(\hat{\beta}-\beta)^T) &= \mathbb{E}((XX^\top)^{-1}X\overrightarrow{\epsilon}\overrightarrow{\epsilon}^\top X^\top(XX^\top)^{-1}) \\ &= (XX^\top)^{-1}X\mathbb{E}(\overrightarrow{\epsilon}\overrightarrow{\epsilon}^\top)X^\top(XX^\top)^{-1} \\ &= \sigma^2(XX^\top)^{-1}.\end{aligned}$$

Now, suppose we can *choose* the columns x_i of X. Assume these can be chosen from $\mathbf{X} = \{x_1, \ldots, x_p\}$. Suppose we choose x_i n_i times ($n_i \geq 0$, $N := \sum_i n_i$); then, X as above is $[\ldots \underbrace{x_i \ldots x_i}_{n_i \text{ times}} \ldots]$.

Let $w_i := n_i/N$, so $w_i \geq 0$ and $e^\top w = 1$. Then we find $\hat{\beta} = (XWX^\top)^{-1}XW\bar{y}$, where $W := \operatorname{diag}(w)$, we have redefined X as $[x_1, \ldots, x_p]$, and $\bar{y}_i$ is the mean of the n_i y_i's corresponding to x_i. Then $\operatorname{cov}(\hat{\beta}) = \frac{\sigma^2}{N}(XWX^\top)^{-1}$. Note we can then choose W to make the covariance smaller while keeping N fixed.

That is, we will choose W to make the determinant of $(XWX^\top)^{-1}$ small, or equivalently $\det(XWX^\top)$ large. This gives

$$\begin{aligned}&\max_w \det(XWX^\top) \\ &\quad e^\top w = 1, \\ &\quad Nw \text{ integer,}\end{aligned}$$

which is a nonlinear integer optimization problem.

We now relax the integer constraints to get the D-optimal design problem:

$$\begin{aligned}&\max_{u \in \mathrm{R}^p} \det(XUX^\top) \\ &\quad e^\top u = 1, \\ &\quad u \geq 0.\end{aligned}$$

If we instead maximize $\ln\det(XUX^\top)$, we have a concave objective function. We'll look at its Lagrangian dual and discuss briefly algorithms next time.