Advanced topics in microeconometrics: Matching Models and their Applications

Alfred Galichon (NYU and ScPo)

Spring 2018

Lecture 7. Convex analysis and nonlinear inverse problems

- Convex analytic notions:
- Inverse problems
- Regularization: entropic, lasso, nuclear norm
- Iterative methods, proximal gradient algorithms

- [OTME], Ch. 6
- Rockafellar (1970). *Convex analysis*. Princeton.

Section 1

Theory

- Assume that P and Q have a convex support with nonempty interior. Recall that if a dual minimizer (u, v) exists, u and v are related by

$$v(y) = \max_{x \in \mathbb{R}^d} \left\{ x^\intercal y - u(x) \right\} \tag{1}$$

$$u(x) = \max_{y \in \mathbb{R}^d} \left\{ x^\intercal y - v(y) \right\} \tag{2}$$

(we can always assign the value $+\infty$ to u outside of the support of P and same for v).

- This expression is a fundamental tool in convex analysis: it is called the *Legendre-Fenchel transform*, which is defined in general by:

DEFINITION

The Legendre-Fenchel transform of u is defined by

$$u^*(y) = \sup_{x \in \mathbb{R}^d} \left\{ x^\intercal y - u(x) \right\}. \tag{3}$$

PROPOSITION

The following holds:
(i) u^* *is convex.*
(ii) $u_1 \leq u_2$ *implies* $u_1^* \geq u_2^*$.
(iii) (Fenchel's inequality): $u(x) + u^*(y) \geq x^\intercal y$.
(iv) $u^{**} \leq u$ *with equality iff* u *is convex.*

As an immediate corollary of (iv), we get the fundamental result:

PROPOSITION

If u *is convex, then* $u = (u^*)^*$. *The converse holds true.*

EXAMPLE

(i) For $u(x) = |x|^2/2$, one gets $u^*(y) = |y|^2/2$.
(ii) For $u(x) = \sum_i \lambda_i x_i^2/2$, $\lambda_i > 0$, one gets $u^*(y) = \sum_i \lambda_i^{-1} y_i^2/2$.
(iii) More generally, for $u(x) = x^\intercal \Sigma x/2$, where Σ is a positive definite matrix, one has $u^*(y) = y^\intercal \Sigma^{-1} y/2$.
(iv) The entropy function

$$u(x) = \begin{cases} \sum_i x_i \ln x_i \text{ for } x \geq 0,\ \sum_i x_i = 1 \\ +\infty \text{ otherwise} \end{cases}$$

has a Legendre transform which is the log-partition function, a.k.a. logit function

$$u^*(y) = \ln\left(\sum_i e^{y_i}\right).$$

(v) Let $p > 1$ and $u(x) = \frac{1}{p}\|x\|^p$, where $\|.\|$ is the Euclidean norm. Then $u^*(y) = \frac{1}{q}\|y\|^q$, where $q > 1$ such that $1/p + 1/q = 1$.

We now restate the demand sets of workers and firms in terms of subdifferentials of convex functions. For this, let us recall the basic economic interpretation of relations (1)-(2), which we had previously spelled out: Expression (1) captures the problem of a firm of type y, which hires a worker x who offers the best trade-off between production if hired by y (that is $\Phi(x, y) = x^{\intercal} y$) and wage $u(x)$. Thus, firm y will be willing to match with any worker whithin the set of maximizers of (1), while worker x will be willing to match with any firm whithin the set of maximizers of (2). The set of maximizers of (1) and of (2) are called *subdifferentials* of v and u,

- The subdifferential is formally defined as follows.

DEFINITION

Let $u : \mathbb{R}^d \to \mathbb{R}$. The subdifferential of u at x, denoted $\partial u(x)$, is the set of $y \in \mathbb{R}^d$ such that $\forall \tilde{x} \in \mathbb{R}^d$, $u(\tilde{x}) \geq u(x) + y^\intercal (\tilde{x} - x)$.

- The definition does *not* require u to be convex; however, if u is convex, Definition 5 immediately implies that

$$\partial u(x) = \arg\max_y \left\{ x^\intercal y - u^*(y) \right\}, \tag{4}$$

hence the subdifferential of a convex function is always nonempty (while the subdifferential of a non-convex function can be empty in general).

It also follows that if u is a convex function, the following statements are equivalent:

$$\text{(i) } u(x) + u^*(y) = x^\intercal y \tag{5}$$
$$\text{(ii) } y \in \partial u(x) \tag{6}$$
$$\text{(iii) } x \in \partial u^*(y). \tag{7}$$

Going back to our worker-firm example, this has a straightforward economic interpretation. If worker x chooses firm y, then y maximizes $x^\intercal \tilde{y} - u^*(\tilde{y})$ over $\tilde{y}$, thus $y \in \partial u(x)$. This means that while worker x's equilibrium wage $u(x)$ is in general greater or equal than the value $x^\intercal y - u^*(y)$ she can extract from firm y, those two values necessarily coincide if x and y are willing to match, in which case $u(x) + u^*(y) = x^\intercal y$.

These considerations allow us to relate the solutions to the primal and dual problems. Recall that in the finite-dimensional case, where the primal and the dual problems are related by a complementary slackness condition. In the present case, let $(X, Y) \sim \pi$ be a solution to the primal problem, and (u, u^*) be a solution to the dual problem. Then almost surely X and Y are willing to match, which, by the previous discussion, implies that

$$u(X) + u^*(Y) = X^\intercal Y, \tag{8}$$

or equivalently $Y \in \partial u(X)$ or in turn $X \in \partial u^*(Y)$. In other words, the support of π is included in the set $\left\{(x, y) : u(x) + u^*(y) = x^\intercal y\right\}$. This condition appears as the correct generalization of the complementary slackness condition in the finite-dimensional case. Without surprise, taking the expectation with respect to π of equality (8) yields the equality between the value of the dual problem on the left-hand side, and the value of the primal problem on the right-hand side.

More can be said when u is differentiable at x. In that case, it is not hard to show that $\partial u\left(x\right)=\left\{\nabla u\left(x\right)\right\}$, i.e. contains only one point, which is $\nabla u\left(x\right)=\left(\partial u\left(x\right)/\partial x_i\right)_i$, the vector of partial derivatives of u, or gradient of u. Similarly, if u^* is differentiable at y, then $\partial u^*\left(y\right)=\left\{\nabla u^*\left(y\right)\right\}$. Hence, if u and v are differentiable, then the equivalence between (6) and (7) implies that $y=\nabla u\left(x\right)$ if and only if $x=\nabla u^*\left(x\right)$, that is

$$\left(\nabla u\right)^{-1}=\nabla u^*. \tag{9}$$

Alternatively, relation (9) can be seen as a duality between first-order conditions and the envelope theorem. First order conditions in the firm's problem (1) implies that if worker x is chosen by firm y, then $\nabla u\left(x\right)=y$, but the envelope theorem implies that the gradient in y of the firm's indirect profit $u^*\left(y\right)$ is given by $\nabla u^*\left(y\right)=x$, where x is chosen by y. Thus the first-order conditions and the envelope theorem are “conjugate” in the sense of convex analysis.

- It's time to make a pause—and take a breath. Thanks to optimal transport, we have seen a natural way to introduce a very useful toolbox, convex analysis, and make sense of u^*, ∂u, ∂u^*, etc. because these objects interpret particularly well using the language of two-sided matching between workers and firms.
- We will need a lot of convex analysis in the sequel of this course. Doing so, we shall leave the interpretation as worker-firms matching, and we will use convex analysis as a mere toolbox.
- The remaining part of this lecture exemplifies this. We shall manipulate convex functions, their Legendre-Fenchel transforms, and their subdifferentials as mathematical objects, and without assigning them an interpretation as payoff functions in a matching problem.

- In the sequel, we shall see an important class of inverse problems called "demand inversion problem". Assume that choosing some alternative j yieds average utility U_j to the consumer. Let s_j be the market share of j, i.e. the probability that the consumer chooses j. Typically s is observed and one seeks to identify U.
- As we shall see, we can often write the model as

  $$s \in \partial G\left(U\right)$$

  where G is a convex function.
- Therefore, the inverse problem amounts to inverting this relationship; thus

  $$U \in \partial G^{*}\left(s\right)$$

  however, the set of U's that rationalize a given vector of market share is potentially large.

- Take the simplest example, where j is chosen if $j \in \arg\max_j \{U_j\}$. This is the revealed preference model, which assumes that all consumers are heterogenous.
- Then one may take $G(U) = \max_j U_j$, so that $\partial G(U)$ is the set of probability vectors s supported on $\arg\max_j U_j$. One has

$$s \in \partial G(U) \iff U \in \partial G^*(s) \iff \begin{cases} s \geq 0, \ \sum_j s_j = 1 \\ s_j > 0 \Rightarrow j \in \arg\max_k \{U_k\} \end{cases}$$

- This is not very useful for econometrics purposes. Indeed, assuming that the market shares are all positive, this means that the only compatible utility vectors that are those such that (U_j)=constant.

- The first motive of regularization arises from the desire to account for unobserved heterogeneity. Start from the unregularized problem $U \in \partial G^* (s)$, which writes

$$s \in \arg\max_{s \geq 0} \left\{ \sum_j s_j U_j : \sum_j s_j = 1 \right\},$$

and insert a penalization $\sigma I(s)$ in the objective function, where $\sigma > 0$ is a parameter, and I is convex, so that the regularized problem is

$$s \in \arg\max_{s \geq 0} \left\{ \sum_j s_j U_j - \sigma I(s) : \sum_j s_j = 1 \right\}.$$

- A particularly popular regularization is the *entropic regularization*, i.e.

$$I(s) = \sum_j s_j \ln s_j$$

in which case one has

$$s_j = \frac{e^{U_j/\sigma}}{\sum_k e^{U_k/\sigma}}$$

which is the logit model. Later on, we shall see a microfoundation this model as a random utility model, but it is helpful to see the logit model as a regularization of the revealed preference model.
- The parameter σ controls the amount of observable heterogeneity we are allowing in the model. When the weight σ decreases to zero, s tends to a particular vector of market shares selected in the set of distribution whose support is in the argmax (randomness decreases); when σ increases, s tends to the uniform distribution (randomness increases).
- In the case of this model (logit model), one has classically

$$\begin{cases} G(U) = \sigma \log \sum_j \exp(U_j/\sigma) \\ G^*(s) = \sigma \sum_j s_j \log s_j. \end{cases}$$

- In some cases, the researcher wants to incorporate beliefs about the structural parameter of interest (here, U). For instance, U may be sparse, i.e. $\#\{j : U_j \neq 0\}$ is small.
- In this case, L1 penalization (Lasso) is a method of choice. Start from the unpenalized logit model, where U is obtained from s by

$$U \in \arg\max_U \left\{ \sum_j s_j U_j - \sigma \log \sum_j \exp\left(U_j/\sigma\right) \right\}$$

and add a penalty $\gamma \left|U\right|_1 = \gamma \sum_j \left|\lambda_j\right|$ to "pull" the solution toward sparse U's. (Note that this time, it is U we are penalizing, not s.)

- The problem becomes

$$U \in \arg\max_U \left\{ \sum_j s_j U_j - \sigma \log \sum_j \exp\left(U_j/\sigma\right) - \gamma \left|U\right|_{L^1} \right\}$$

and unlike the entropic regularization, the penalization is nonsmooth. Fortunately, there are very powerful methods to handle this: proximal gradient algorithms.

- To compute
$$\min f(x) + \gamma |x|_1$$
we use the proximal gradient algorithm:
$$x^{t+1} = prox_\epsilon \left(x^t - \epsilon \nabla f\left(x^t\right)\right)$$
where
$$prox_\epsilon (z)_i = (z_i - \epsilon) 1 \{z_i \geq \epsilon\} + (z_i + \epsilon) 1 \{z_i \leq -\epsilon\}.$$
- Intuition: x^{t+1} minimizes $\gamma |x|_1 + \frac{1}{2\epsilon} \|x - x^t + \epsilon \nabla f(x^t)\|_2^2$, which the original function where f has been replaced by a quadratic approximation.

Section 2

Coding

- See Keith's presentation slides.