IE 495 – Lecture 5

Stochastic Programming – Math Review and MultiPeriod Models

Prof. Jeff Linderoth

January 27, 2003

Outline

- Homework – questions?
  - I would start on it fairly soon if I were you...
- A fairly lengthy math (review?) session
  - Differentiability
  - KKT Conditions
- Modeling Examples
  - Jacob and MIT
  - “Multi-period” production planning

Yucky Math Review – Derivative

- Let f be a function from $\Re^n \mapsto \Re$. The *directional derivative* f' of f with respect to the direction d is

$$f'(x, d) = \lim_{\lambda \to 0} \frac{f(x + \lambda d) - f(x)}{\lambda}$$

- If this direction derivative exists and has the same value for all $d \in \Re^n$, then f is *differentiable.*
- The unique value of the derivative is called the *gradient* of f at x
  - ◇ We denote its value as $\nabla f(x)$.

Not Everything is Differentiable

- Probably, everything you have ever tried to optimize has been differentiable.

⋆ This will not be the case in this class!

- Even nice, simple, convex functions may not be differentiable at all points in their domain.
  - ◇ Examples?
- A vector $\eta \in \Re^n$ is a *subgradient* of a convex function f at a point x iff (if and only if)
  - ◇ $f(z) \geq f(x) + \eta^T(z - x) \qquad \forall z \in \Re^n$
  - ◇ The graph of the (linear) function $h(z) = f(x) + \eta^T(z - x)$ is a supporting hyperplane to the convex set $epi(f)$ at the point $(x, f(x))$.

More Definitions

- The set of all subgradients of f at x is called the *subdifferential* of f at x.
  - ◇ Denoted by $\partial f(x)$

? Is $\partial f(x)$ a convex set?

- Thm: $\eta \in \partial f(x)$ iff
  - ◇ $f'(x, d) \geq \eta^T d \qquad \forall d \in \Re^n$

Optimality Conditions

- We are interested in determining conditions under which we can verify that a solution is optimal.
- To KISS, we will (for now) focus on *minimizing* functions that are
  - ◇ One-dimensional
  - ◇ Continuous ($|f(a) - f(b)| \leq L|a - b|$)
  - ◇ Differentiable
- Recall: a function $f(x)$ is convex on a set S if for all $a \in S$ and $b \in S$, $f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)b$.

Why do we care?

- Because they are important
- Because Prof. Linderoth says so!
- Many optimization algorithms work to find points that satisfy these conditions
- When faced with a problem that you don't know how to handle, write down the optimality conditions
- Often you can learn a lot about a problem, by examining the properties of its optimal solutions.

Preliminaries

Call the following problem P:

$$z^* = \min f(x) : x \in S$$

- Def: Any point $x^* \in S$ that gives a value of $f(x^*) = z^*$ is the *global minimum* of P.

  ◇

  $x^* = \arg\min_{x \in S} f(x)$.

- Def: *Local minimum* of P: Any point $x^l \in S$ such that $f(x^l) \geq f(y)$ for all y "in the neighborhood" of x^l. ($y \in S \cap N_\epsilon(x^l)$).

- Thm: Assume S is convex, then if $f(x)$ is convex on S, then any local minimum of P is a global minimum of P.

Oh No – A Proof!

- Since x^l is a local minimum, $\exists N_\epsilon(x^l)$ around x^l such that
  - $f(x) \geq f(x^l) \; \forall \; x \in S \cap N_\epsilon(x^l)$.
- Suppose that x^l is *not* a global minimum, so $\exists \; \hat{x} \in S$ such that $f(\hat{x}) < f(x^l)$.
- Since f is convex, $\forall \lambda \in [0, 1]$,
  - $f(\lambda\hat{x} + (1 - \lambda)x^l) \leq \lambda f(\hat{x}) + (1 - \lambda)f(x^l) < \lambda f(x^l) + (1 - \lambda)f(x^l) = f(x^l)$
- For $\lambda > 0$ and very small $\lambda\hat{x} + (1 - \lambda)x^l \in S \cap N_\epsilon(x^l)$.
- But this contradicts $f(x) \geq f(x^l) \; \forall \; x \in S \cap N_\epsilon(x^l)$. Q. E. D.

Starting Simple – Optimizing 1-D functions

Consider optimizing the following function (for a scalar variable $x \in \Re^1$):

$$z^* = \min f(x)$$

Call an optimal solution to this problem x^*. ($x^* = \arg\min f(x)$).

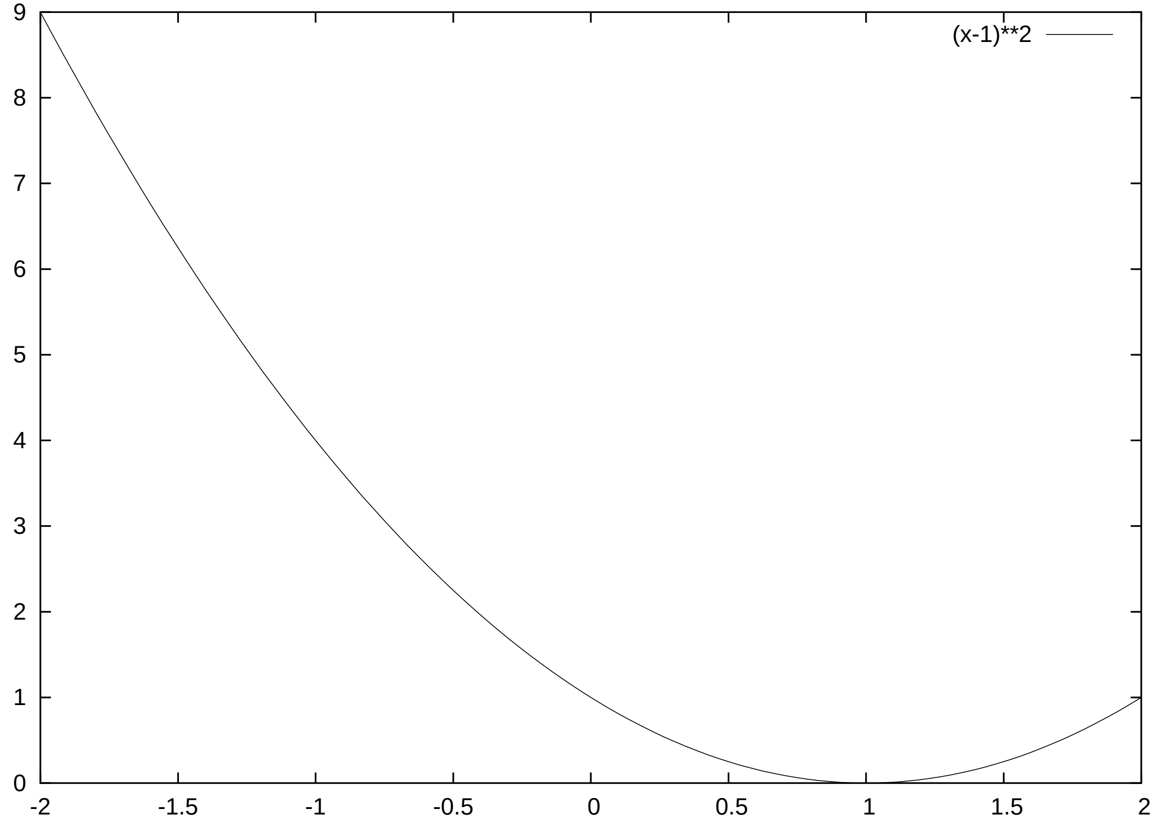
What is *necessary* for a point x to be an optimal solution?

- ⋆ $f'(x) = 0$

Ex. $f(x) = (x-1)^2$

- ◇ $f'(x) = 2(x-1) = 0 \Leftrightarrow x = 1$

$$f(x) = (x-1)^2$$



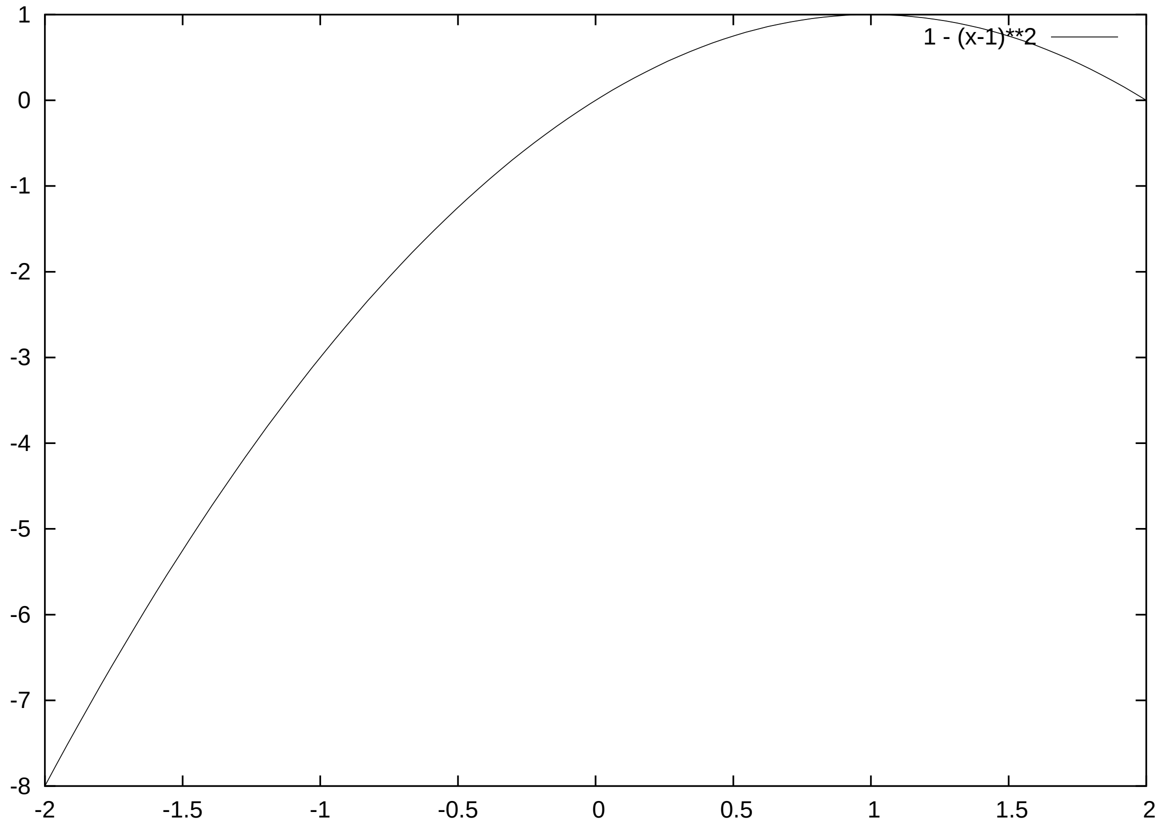
Is That All We Need?

- Is $f'(x) = 0$ also *sufficient* for x to be a (locally) optimal solution?

Ex. $f(x) = 1 - (x-1)^2$

- ◇ $f'(x) = -2(x-1) = 0 \Leftrightarrow x = 1$

$$f(x) = 1 - (x-1)^2$$



Obviously Not

- Since $x = 1$ is a local minimum of $f(x) = 1 - (x-1)^2$, the $f'(x) = 0$ condition is obviously not all we need to ensure that we get a local minimum

? What *is* the *sufficient* condition for a point $\hat{x}$ to be (locally) optimal?

⇒ $f''(\hat{x}) > 0$!

  ◇ This is equivalent to saying that $f(x)$ is *convex* at $\hat{x}$.

  ? Who has heard of the following terms?

    ◇ "Hessian Matrix"?

    ◇ "Positive (Semi)-definite"?

- If $f(x)$ is convex for all x, then (from the previous Thm.) any local minimum is also a global minimum.

(1-D) Constrained Optimization

Now we consider the following problem for scalar variable $x \in \Re^1$.

$$z^* = \min_{0 \leq x \leq u} f(x)$$

- There are three cases for where an optimal solution might be
  - $x = 0$
  - $0 < x < u$
  - $x = u$

Breaking it down

- If $0 < x < u$, then the necessary and sufficient conditions for optimality are the *same* as the unconstrained case
- If $x = 0$, then we need $f'(x) \geq 0$ (necessary), $f'' > 0$ (sufficient)
- If $x = u$, then we need $f'(x) \leq 0$ (necessary), $f'' > 0$ (sufficient)

KKT Conditions

- How do these conditions generalize to optimization problems with more than one variable?
- The intuition — if a constraint holds with equality (is binding), then the gradient of the objective function must be pointing in a way that would improve the objective.
- Formally — The negative gradient of the objective function must be a linear combination of the gradients of the binding constraints.
- The "KKT" stands for Karush-Kuhn-Tucker.
  - ◇ Story Time!

⋆ Remember the "Optimality Conditions" from linear programming? These are just the KKT conditions!

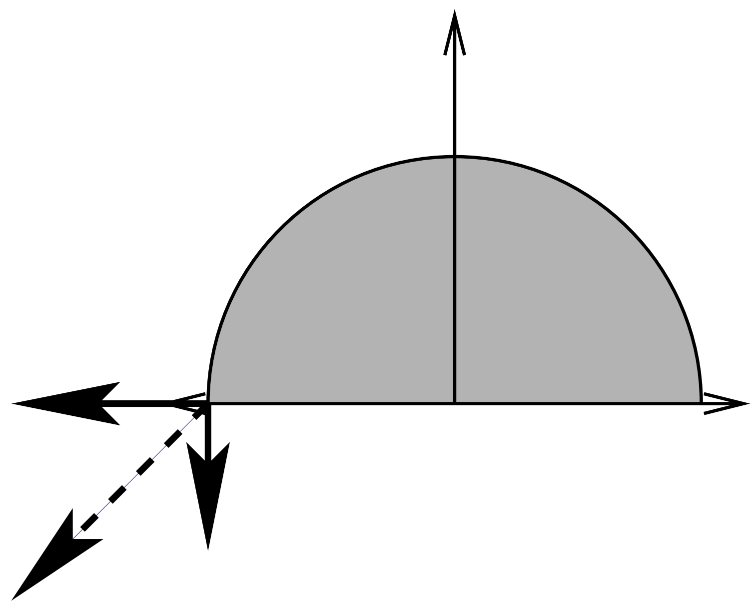
Example ($x \in \Re^2$)

minimize

$$x_1 + x_2$$

subject to

$$\begin{aligned} x_1^2 + x_2^2 &\leq 2 \\ -x_2 &\leq 0 \end{aligned}$$



- You see at the optimal solution $x = (-\sqrt{2}, 0)$,

The Canonical Problem

minimize

$$f(x)$$

subject to

$$\begin{aligned} g_1(x) &\leq b_1 \\ g_2(x) &\leq b_2 \\ &\vdots \\ g_m(x) &\leq b_m \\ -x_1 &\leq 0 \\ -x_2 &\leq 0 \\ &\vdots \\ -x_n &\leq 0 \end{aligned}$$

KKT Conditions

- Geometrically, if $(\hat{x})$ is an optimal solution, then we must be able to write $-\nabla f(\hat{x})$ as a nonnegative linear combination of the binding constraints.
- If a constraint is not binding, it's "weight" must be 0.

$$\begin{aligned} \nabla - f(\hat{x}) &= \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) - \mu \\ \lambda_i = 0 \quad &\text{if} \quad g_i(\hat{x}) < b \qquad (\forall i) \\ \mu_j = 0 \quad &\text{if} \quad \hat{x}_j > 0 \qquad (\forall j) \end{aligned}$$

KKT Conditions

If $\hat{x}$ is an optimal solution to P, then there exists multipliers $\lambda_1, \lambda_2, \ldots, \lambda_m, \mu_1, \mu_2, \ldots, \mu_n$ that satisfy the following conditions:

$$
\begin{aligned}
g_i(\hat{x}) &\leq b_i & \forall i = 1, 2, \ldots, m \\
-x_i &\leq 0 & \forall j = 1, 2, \ldots, n \\
-\frac{\partial f(\hat{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \frac{\partial g(\hat{x})}{\partial x_j} + \mu_j &= 0 & \forall j = 1, 2, \ldots n \\
\lambda_i &\geq 0 & \forall i = 1, 2, \ldots, m \\
\mu_j &\geq 0 & \forall j = 1, 2, \ldots, n \\
\lambda_i(b_i - g_i(\hat{x})) &= 0 & \forall i = 1, 2, \ldots, m \\
\mu_j x_j &= 0 & \forall j = 1, 2, \ldots n
\end{aligned}
$$

Returning to example

minimize

$$x_1 + x_2$$

subject to

$$\begin{aligned} x_1^2 + x_2^2 &\leq 2 \qquad (\lambda) \\ -x_2 &\leq 0 \qquad (\mu) \end{aligned}$$

KKT Conditions

Primal Feasible:

$$\begin{aligned} x_1^2 + x_2^2 &\leq 2 \\ -x_2 &\leq 0 \end{aligned}$$

Dual Feasible:

$$\begin{aligned} \lambda &\geq 0 \\ \mu &\geq 0 \\ -1 &= \lambda(2x_1) \\ -1 &= \lambda(2x_2) - \mu \end{aligned}$$

KKT Conditions, cont.

Complementary Slackness:

$$\begin{aligned} \lambda(2 - x_1^2 - x_2^2) &= 0 \\ \mu x_2 &= 0 \end{aligned}$$

Generalizing to Nondifferentiable Functions

- In full generality, this would require some fairly heavy duty convex analysis.
  - ◇ Convex analysis is a great subject, you should all study it!
- Instead, I first want to show that when passing to nondifferentiable functions, we would replace $\nabla f(x) = 0$ with $0 \in \partial f(x)$.
- I am sorry for all the theorems, but all little more math never *hurt* anyone. (At least as far as I know).

Theorem

- Let $f : \Re^n \mapsto \Re$ be a convex function and let S be a nonempty convex set. $\hat{x} = \arg\min_{x \in S} f(x)$ if (and only if) η is a subgradient of f at $\hat{x}$ such that $\eta^T(x - \hat{x}) \geq 0 \; \forall x \in S$

Proof. (Duh!)

- I will prove only the very, very easy direction.
- If η is a subgradient of f at $\hat{x}$ such that $\eta^T(x - \hat{x}) \geq 0 \; \forall x \in S$,
  - $f(x) \geq f(\hat{x}) + \eta^T(x - \hat{x}) \geq f(\hat{x}) \qquad \forall x \in S.$
  - So $\hat{x} = \arg\min_{x \in S} f(x)$ Q.E.D.

Theorem

- Let $f : \Re^n \mapsto \Re$ be a convex function. $\hat{x} = \arg\min_{x \in \Re^n} f(x)$ if (and only if) $0 \in \partial f(\hat{x})$.

Proof.

- $\hat{x} = \arg\min_{x \in \Re^n} f(x)$ if and only if ($\Leftrightarrow$) η is a subgradient where $\eta^T(x - \hat{x}) \geq 0 \ \forall x \in \Re^n$.
- Choose $x = \hat{x} - \eta$.
- $\eta^T(\hat{x} - \eta - \hat{x}) = -\eta^T \eta \geq 0$.
- This can only happen when $\eta = 0$
  - ◇ $(-\sum \eta_i^2 = 0 \Leftrightarrow \eta_i = 0 \quad \forall i)$.
- So $\eta = 0 \in \partial f(\hat{x})$ Q.E.D.

Now in Full Generality

- **Thm:** For a convex function $f : \Re^n \mapsto \Re$, and convex functions $g_i : Re^n \mapsto \Re, i = 1, 2, \ldots m$, if we have some nice "regularity conditions" (which you should assume we have unless I tell you otherwise), $\hat{x}$ is an optimal solution to $\min\{f(x) : g_i(x) \leq 0 \ \forall i = 1, 2, \ldots m\}$ if and only if the following conditions hold:
  - ◇ $g_i(x) \leq 0 \forall i = 1, 2, \ldots m$
  - ◇ $\exists \lambda_1, \lambda_2, \ldots \lambda_m \in \Re$ such that
    - $0 \in \partial f(\hat{x}) + \sum_{i=1}^m \lambda_i \partial g_i(\hat{x})$.
    - $\lambda_i \geq 0 \ \forall i = 1, 2, \ldots m$
    - $\lambda_i g_i(\hat{x}) = 0 \ \forall i = 1, 2, \ldots m$

Daddy Has Big Plans

- MIT costs \$39,060/year right now
- In 10 years, when Jacob is ready for MIT, it will cost > \$80000/year. (YIKES!)
- Let's design a stochastic programming problem to help us out.
- In Y years, we would like to reach a tuition goal of G.
- We will assume that Helen and I rebalance our portfolio every v years, so that there are $T = Y/v$ times when we need to make a decision about what to buy.
  - ◇ There are T periods in our stochastic programming problem.

Details

- We are given a universe N of investment decisions
- We have a set $\mathcal{T} = \{1, 2, \ldots T\}$ of investment periods
- Let $\omega_{it}, i \in N, t \in \mathcal{T}$ be the return of investment $i \in N$ in period $t \in \mathcal{T}$.
- If we exceed our goal G, we get an interest rate of q that Helen and I can enjoy in our golden years
- If we don't meet the goal of G, Helen and I will have to borrow money at a rate of r so that Jacob can go to MIT.
- We have \$$b$ now.

Variables

- $x_{it}, i \in N, t \in \mathcal{T}$: Amount of money to invest in vehicle i during period t
- y : Excess money at the end of horizon
- w : Shortage in money at the end of the horizon

(Deterministic) Formulation

maximize

$$qy + rw$$

subject to

$$\begin{aligned}
\sum_{i \in N} x_{i1} &= b \\
\sum_{i \in N} \omega_{it} x_{i,t-1} &= \sum_{i \in N} x_{it} \qquad \forall t \in \mathcal{T} \setminus 1 \\
\sum_{i \in N} \omega_{iT} x_{iT} - y + w &= G \\
x_{it} &\geq 0 \qquad \forall i \in N, t \in \mathcal{T} \\
y, w &\geq 0
\end{aligned}$$

Random returns

- As evidenced by our recent performance, my wife and I are bad at picking stocks.
  - ◇ In our defense, returns on investments are *random* variables.
- Imagine that for each there are a number of potential outcomes R for the returns at each time t.

Scenarios

- The scenarios consist of all possible sequences of outcomes.

Ex. Imagine $R = 4$ and $T = 3$. The the scenarios would be...

$t = 1$	$t = 2$	$t = 3$
1	1	1
1	1	2
1	1	3
1	1	4
1	2	1
	$\vdots$	
4	4	4

Making it Stochastic

- $x_{its}, i \in N, t \in \mathcal{T}, s \in S$: Amount of money to invest in vehicle i during period t in scenario s
- y_s : Excess money at the end of horizon in scenario s
- w_s : Shortage in money at the end of the horizon in scenario s

⋆ Note that the (random) return ω_{it} now is like a function of the scenario s.

◇ It depends on the mapping of the scenarios to the scenario tree.

A Stochastic Version

maximize

$$qy_s + rw_s$$

subject to

$$\begin{aligned}
\sum_{i \in N} x_{i1} &= b \\
\sum_{i \in N} \omega_{its} x_{i,t-1,s} &= \sum_{i \in N} x_{its} \qquad \forall t \in \mathcal{T} \setminus 1, \forall s \in S \\
\sum_{i \in N} \omega_{iT} x_{iTs} - y_s + w_s &= G \qquad \forall s \in S \\
x_{its} &\geq 0 \qquad \forall i \in N, t \in \mathcal{T}, \forall s \in S \\
y_s, w_s &\geq 0 \qquad \forall s \in S
\end{aligned}$$

Next time

? Is this correct?

- Answer the question above...
- Writing the deterministic equivalent of multistage problems
- (Maybe) one more modeling example
- Properties of the recourse function. (Starting BL 3.1)