Applications of Convexity, JHU, March 30-31, 2006

L. A. Shepp, Rutgers University

e-mail: shepp@stat.rutgers.edu

It's a pleasure to be back at Hopkins, where my son, Eric, spent 5 years getting two degrees and a fine education. Eric is now a hydrologist/engineer in Tucson. He came to one of my lectures here and had some difficulty with the math, but was pleased to note that my math wasn't so hot either, since my estimate of the size of my audience was seriously in error. This became a family joke.

Thanks to my old friends including Jim Fill for this invitation. I hope this topic will appeal broadly because convexity is so widely used in mathematics often without the practitioners realizing that they are dealing with convexity. I will illustrate with examples.

Here is one such problem: You owe the mafia \$1 million and you have only \$0.2 million. You are in a casino and can stake any amount s of money, at any time, but s cannot exceed your current fortune f, i.e., $s \leq f$ at any one of the games available. The game on table T_r gives odds r. That is if you bet s at odds r and win then your new fortune is $f + rs$, and this occurs with probability w. If you bet s and lose then your new fortune is $f - s$. It is natural to assume that $w = \frac{1+c}{1+r}$ where $c \in (-1, 0]$ because this makes the expected net return on a one-dollar bet at odds r the same value, $c \leq 0$, on every table. Thus all tables are subfair with the same c. The problem is to choose the odds r and the stake s so as to maximize the probability that you will reach fortune one and remain alive. You start with fortune $f_0 = .2$ and bet some amount which you can choose at some odds which you can choose and move to fortune f_1 then to f_2, etc. You stop when $f_n = 0$ (and you die) or when $f_n = 1$, and you give up gambling and return to your family and live happily ever after.

If there is only one table, say with even money odds, $r = 1$, and say $w = .49$, to make it subfair. Note that if $w = .51$ then you live with probability one by making very small bets and exploiting the law of large numbers. But in the subfair or Dubins casino, this is an old problem invented and completely solved by Lester Dubins and Jimmie Savage and discussed in their book, "How to Gamble if You Must". They guessed correctly that one cannot do better than to *play boldly*, i.e. set $s = f$ or $s = 1 - f$, whichever is smaller. They showed that if f_0 has the binary expansion

$$f_0 = \sum_{k=0}^{\infty} 2^{-n_k}, 1 \leq n_0 < n_1 < \ldots,$$

then the optimal survival probability attained by bold play is

$$Q(f_0) = \sum_{k=0}^{\infty} w^{n_k - k}(1 - w)^k$$

which is a continuous, singular function in that it increases from $Q(0) = 0$ to $Q(1) = 1$ on a set of measure zero. This is because it is clear that

$Q(f) = wQ(2f), f \leq \frac{1}{2}$, and

$Q(f) = w + Q(2f - 1), \frac{1}{2} < f < 1$,

and so $Q(\frac{1}{2}) = w$, $Q(\frac{1}{4}) = w^2$, and e.g.

$Q(\frac{1}{2}+\frac{1}{2^n}) = w+(1-w)Q(\frac{2}{2^n}) = w+(1-w)w^{n-1}$.

Now a careful induction argument shows that for any $s \leq f$, we have

$Q(f + s)w + Q(f - s)(1 - w) \leq Q(f)$.

and it follows from this that for any strategy of choosing stakes s, the process $Q(f_n)$ is a supermartingale. The supermartingale inequality which is due to Doob now gives that $EQ(f_\infty) \leq Q(f_0)$. Since $f_\infty = 1$ iff one gets to $f_n = 1$ and $f_\infty = 0$ otherwise, it follows that $P(f_\infty = 1) \leq Q(f_0)$ for any staking plan.

Since bold play is a staking plan, it follows that no staking plan can beat bold play. Remarkably however, there are other plans that do as well as bold play. Can Dubins's theorem be proved without using submartingales? It seems so intuitive that many would say yes based on *minimizing the number of bets* The answer is no, though many alleged proofs have been published. Most of these "proofs" prove too much, that bold play is uniquely optimal, which is false. One **must** use supermartingales because one needs to use the fact that when a bet is made, the outcome is not known. There is no way that anyone knows how to make use of this except for supermartingales. Dubins never uses the word supermartingale in his book for some unknown reason, but his proof uses the supermartingale idea.

The same result holds for any odds r if there is only one table in the Dubins casino. Moreover, suppose one has to pay interest after every bet so one moves fro f to $\frac{f \pm s}{1+a}$ where $a > 0$. Now it's even more "obvious" that one should play boldly - the only problem is that it is false (!), except for certain initial fortunes f_0, eg when you can reach fortune one by bold play in one or more winning steps.

But what if we are in the Vardi casino (suggested by Yehuda Vardi before his untimely death), where one can play on *any* table and for each odds r there is a table, T_r.

What is the best way to make bets s and choose tables T_r so as to maximize the probability to remain alive? That is, what is

$$V(f) = \sup_{\{\mathcal{S},\mathcal{R}\}} \mathbf{P}^{\mathcal{S},\mathcal{R}}[F_\infty = 1 | F_0 = f]?$$

Here the supremum is over all strategies $\mathcal{S}$ of choosing bets and all strategies $\mathcal{R}$ of choosing odds.

Besides thinking about what the best strategy is, you may be wondering why this is a problem in convexity! I will try to show you that it is.

The answer is neat: $V(f) = 1-(1-f)^{1+c}$; but how to achieve it is not so simple.

In fact it can't be achieved! It's the supremum over all staking plans but it is not attained. Here is how to come within ϵ. If $f \leq \epsilon$, set $s = f$. Otherwise set $s = \epsilon\frac{1-f}{1-\epsilon}$. Always bet on that table where you reach fortune one in one shot if you win. This gives a calculable probability to reach fortune one and it is increasing as ϵ decreases to zero to $V(f) = 1 - (1 - f)^{1+c}$. This is not hard to show. It is also not hard to show that $V(f_n)$ is a supermartingale for any plan $\mathcal{R}, \mathcal{S}$ so the same proof works to show that this answer Dubins argument applies. Note that $V(f)$ gives an upper bound for every casino where a bet of a dollar on any table gives expected return $\leq c$, so it's quite useful. Note that we had to guess the approximate optimal strategy above, and it is a compromise strategy between bold play (in r) and making small bets (in s, as if it were a superfair game).

On one level one can see that the key idea is martingale theory, but if one steps back a little more one sees that the real idea is convexity. In linear programming, or convexity theory, one maximizes a linear functional under linear inequalities. Linear inequalities always define a convex set, and the linear functional achieves its maximum at an extreme point of the convex set. Interior point and exterior point algorithms for linear programming really find the extreme point of the convex set. Here we found it by guessing the winning strategy. In concompact situations like the Vardi casino, the supremum may not be ttained but the same ideas of convexity work. But where is the linear functional in the Dubins and Vardi problems? It is the choice of Q or V which makes $Q(f_n)$ a supermartingale. Any such Q will give an upper pound for the problem of maximizing the probability to reach fortune one by the supermartingale argument, and the set of all such Q or V is a comvex set. There is an optimal extreme point Q or an ϵ-optiml V, also an extreme point of the convex set. I hope it is clear that all of stochastic optimization theory is just convexity or (infinite dimensional) linear programming. It's useful to know this, but you still have to use your intuition to get the answer in each particular case, This is why I love this subject - it seems hard to program a computer to develop the intuition to come up with bold play or with the ϵ strategy above so human beings will be needed for a while to get answers to optimal control problems.

Convexity also arises in tomography, which may seem surprising. Tomography is based on Radon's theorem that a function $f(x, y)$ in L_1 can be uniquely (up to a set of zero measure) reconstructed from its Radon transform or line integrals, for all straight lines, L,

$$P_f(L) = \int_L f ds.$$

In practical applications, like CAT scanners, one measures $P_f(L)$ using X-rays, to arbitrary precision, but of course for only a fixed set, $\mathcal{L}$

of lines L. The set of g's with given line integrals, $P_g(L) = P_f(L)$, for all $L \in \mathcal{L}$ is clearly a convex set, C_f. The extreme points of C_f define the "error bars" of the CAT scan. To my knowledge nobody has really explored the extreme points of this set for practical cases of $\mathcal{L}$. Of course we know that CAT scanning is effective, but what are the true error bars?

Note that these are the error bars for noiseless measurements. But it is not unreasonable to assume in CAT scanning that there is no noise because that can effectively be achieved. Let $\mathcal{L}$ be the set of all lines with any one of n given slopes. This is an uncountable set, but we are mathematicians.

Theorem: Given any $0 < f < 1$ there exists $g = 0, 1$ only, with the same line integrals for all lines $L \in \mathcal{L}$.

Proof. Consider an extreme $g \in C_f$. It can be shown from weak compactness of C_f that an extreme g must be zero or one a.e. qed.

The conclusion is that there are no error bars on the value of $f(x, y)$ at a fixed point. However it is more interesting to know the error bars on the integral of f over a pixel or volel, i.e. error bars on $\int fh$ where $h(x, y)$ is an approximate delat function. This has not been done and it seems not that hard to do. I have tried to show you that tomography can be viewed also as a chapter in convexity theory.

In my lecture tomorrow I will try to convince you that convexity is the way that Hardy, Littlewood and Polya should have proved the inequalities in their first chapter. They of course do use convexity, but not so systematically, perhaps because von Neumann came along later on and after him, convexity became better understood. I will give examples of how to use convexity to give new proofs of Schwartz's and other inequalities. Of course there are limits to the method and some inequalities cannot be handled by convexity alone. I will give an example of one of these also still not understood.

Problems in Convexity

Convexity arises in mathematics in many contexts, and in particular in stochastic optimization theory and linear programming. These two are actually the same subject, despite the fact that many people in stochastic optimization (including myself for many years) seem not fully aware of the link between the two subjects, namely, duality theory. One can also view tomography in terms of convexity theory. Also, many of the standard inequalities in Hardy, Littlewood, and Polya, which was written before von Neumann introduced linear programming, can be proved by duality theory in convexity.

For a positive r.v. X, let $\mu_\alpha(X)$ be the αth moment of X. Mark Brown's inequality [Brown] states that for positive and independent r.v.'s, X, Y, $F(X+Y) \geq F(X) + F(Y)$, where $F(X)$ is the ratio μ_{-1} over μ_{-2}. We prove this inequality by a method which can be used to prove the Schwarz inequality but which is not widely appreciated. We use the same method to prove and to generalize Brown's inequality.

The Schwarz inequality is better known and perhaps more useful than the Brown inequality but they are similar. The Schwarz inequality can be stated for a general measure space but easily reduces to the statement that

$$EX^2EY^2 \geq (EXY)^2,$$

where X and Y are any r.v.'s on a common probability space. Equality holds if and only if X and Y are proportional.

The Brown inequality refers to *positive and independent* r.v.'s on a common probability space and it asserts that

$$\frac{E\frac{1}{X+Y}}{E\frac{1}{(X+Y})^2} \geq \frac{E\frac{1}{X}}{E\frac{1}{X^2}} + \frac{E\frac{1}{Y}}{E\frac{1}{Y^2}}.$$

Equality holds if and only if both X and Y are constants.

The Schwarz inequality can be proved by constructing the product probability space with the product measure so that X_1, Y_1 and X_2, Y_2 are two *independent* pairs of r.v.'s on the product space each with the joint distribution of X, Y. Note that for any four numbers, x_1, y_1, x_2, y_2, the homogeneous polynomial

$$x_1^2 y_2^2 + x_2^2 y_1^2 - 2x_1 y_1 x_2 y_2 = (x_1 y_2 - y_1 x_2)^2 \geq 0.$$

Now substitute X_i, Y_i for $x_i, y_i, i = 1, 2$ and take expectations, using the independence of r.v.'s with different subscripts to obtain that

$$2EX^2 EY^2 \geq 2(EXY)^2$$

and the proof is complete after dividing by 2. The only case of equality is when $X_1 Y_2 - X_2 Y_1 \equiv 0$, that is when the ratios $\frac{X_i}{Y_i}$ are constant since they are independent for $i = 1, 2$.

The Brown inequality is next proved by the same method, but the proof is a bit trickier. We “clear of fractions” by multiplying by $EX^{-2}EY^{-2}E(X+Y)^{-2}$, and so it is equivalent to show that

$$E\frac{1}{X+Y}EX^{-2}EY^{-2} - E\frac{1}{(X+Y)^2}(EX^{-1}EY^{-2}+$$

$$+EX^{-2}EY^{-1}) \geq 0.$$

To use the method used above for the Schwarz inequality, we again construct the product probability space on which two independent pairs of independent r.v.'s X_i, Y_i are defined. Then *if* we could show that for any four numbers x_1, y_1, x_2, y_2 the function

$$f(x_1, y_1, x_2, y_2) \equiv \frac{1}{x_1+y_1}\frac{1}{x_2^2}\frac{1}{y_2^2} - \frac{1}{(x_1+y_1)^2}(x_2^{-1}y_2^{-2}+ x_1^{-2}y_2^{-1})$$

is everywhere nonnegative, then it would easily follow that

$$Ef(X_1, Y_1, X_2, Y_2) \geq 0$$

which would then prove the Brown inequality, but (alas) f takes negative values. Alternatively, *if* we could show that $f(x_1, y_1, x_2, y_2) + f(x_2, y_2, x_1, y_1)$ is everywhere nonnegative, then the same proof would give the Brown inequality because upon substituting r.v.'s X_i, Y_i for x_i, y_i we would get the desired inequality after dividing by two. Again (alas), there are numbers $x_i, y_i, i = 1, 2$ for which this form is also negative. Fortunately, we have one last chance. If we can show that the doubly mixed (symmetric in x_1, x_2 and in y_1, y_2) form

$$f(x_1, y_1, x_2, y_2) + f(x_1, y_2, x_2, y_1) + f(x_2, y_1, x_1, y_2) + f(x_2, y_2, x_1, y_1) \geq 0$$

for all positive values of $x_i, y_i, i = 1, 2$, then substituting X_i, Y_i for x_i, y_i, taking expecations and using the independence of X_1, X_2, Y_1, Y_2 we get the Brown inequality after dividing by 4.

The proof is not really all that hard. We define

$$g(x_1,y_1,x_2,y_2) = f(x_1,y_1,x_2,y_2)+f(x_1,y_2,x_2,y_1)+ f(x_2,y_1,x_1,y_2) + f(x_2,y_2,x_1,y_1)$$

and prove that $g \geq 0$ by exhibiting it as the sum of two positive things.

Indeed, we can write

$$g(x_1,y_1,x_2,y_2) = \frac{x_1+y_1-x_2-y_2}{(x_1+y_1)^2x_2^2y_2^2} + \frac{x_2+y_1-x_1-y_2}{(x_1+y_1)^2x_1^2y_2^2} + \frac{x_1+y_2-x_2-y_1}{(x_1+y_2)^2x_2^2y_1^2} + \frac{x_2+y_2-x_1-y_1}{(x_2+y_2)^2x_1^2y_1^2}.$$

Gathering terms with the factor $x_1 - x_2$ and $y_1 - y_2$, we can write

$$g(x_1,y_1,x_2,y_2) = (x_1 - x_2)\Big[\frac{1}{(x_1+y_1)^2x_2^2y_2^2} - \frac{1}{(x_2+y_1)^2x_1^2y_2^2} + \frac{1}{(x_1+y_2)^2x_2^2y_1^2} - \frac{1}{(x_2+y_2)^2x_1^2y_1^2}\Big] +$$

$$+(y_1-y_2)\Big[\frac{1}{(x_1+y_1)^2x_2^2y_2^2} - \frac{1}{(x_1+y_2)^2x_2^2y_1^2} + \frac{1}{(x_2+y_1)^2x_1^2y_2^2} - \frac{1}{(x_2+y_2)^2x_1^2y_1^2}\Big].$$

Now we can factor out another $x_1 - x_2$ and $y_1 - y_2$ respectively to write

$$g(x_1, y_1, x_2, y_2) =$$

$$(x_1 - x_2)^2 [\frac{1}{y_2^2} \frac{y_1((x_2+y_1)x_1+(x_1+y_1)x_2)}{(x_1+y_1)^2(x_2+y_1)^2 x_1^2 x_2^2} +$$

$$+ \frac{1}{y_1^2} \frac{y_2((x_2+y_2)x_1+(x_1+y_2)x_2)}{(x_1+y_2)^2(x_2+y_2)^2 x_1^2 x_2^2}] +$$

$$+ (y_1 - y_2)^2 [\frac{1}{x_2^2} \frac{x_1((x_1+y_2)y_1+(x_1+y_1)y_2)}{(x_1+y_1)^2(x_1+y_2)^2 y_1^2 y_2^2} +$$

$$+ \frac{1}{x_1^2} \frac{x_2((x_2+y_2)y_1+(x_2+y_1)y_2)}{(x_2+y_2)^2(x_2+y_1)^2 y_1^2 y_2^2}].$$

This "bit" of algebra proves the Brown inequality. Since $g(x_1, y_1, x_2, y_2) = 0$ only when $x_1 = x_2$ and $y_1 = y_2$ it follows that equality holds in the Brown inequality if and only if the r.v.'s X and Y are both constant. It is easy to give examples of non-independent X, Y where the Brown inequality fails so that the assumption of independence, used in the proof, is also necessary.

Taglamitsky's theorem: If

$$|f^{(k)}(x)| \leq e^{-x} \text{ for all } x \geq 0, k = 0, 1, \ldots$$

show that $f(x) \equiv ce^{-x}$ for some $|c| \leq 1$.

I have not yet seen Tagamlitsky's proof (1946, Dokl.), but probably it is as follows which came from ideas of Cun-Hui Zhang. Let $g(x) = e^{-x} - f(x)$ and note that $(-1)^n g^{(n)}(x) \geq 0$ for all x and $n = 0, 1, \ldots$. It follows that g is completely monotonic and so is a Laplace trnasform of a positive measure, $g(x) = \int_0^\infty e^{-xt} m(dt)$. Now it is easy to argue that the given inequalities cannot hold unless

$$m([0, 1) \cup (1, \infty)) = 0,$$

and Tagamlitsky's theorem follows. The stated inequalities define a convex set of f's and the above proof can be turned into a convexity argument.

The commonality of the various special problems is not fully understood even by the speaker. A brand-new problem is still partly open: This is the problem (formulated by Jack Denny for a problem in medical imaging) of maximizing

$$P(X + Y > 1)$$

over all independent symmetric random variables X and Y each with variance σ^2. The audience is encouraged to contribute to the solution as well as to the general understanding of the commonality of such problems. Problem solvers (I'm one) often fail to see the forest for the trees.