The Witsenhausen Counterexample: A Hierarchical Search Approach for Nonconvex Optimization Problems

Jonathan T. Lee, Edward Lau, and Yu-Chi (Larry) Ho, *Life Fellow, IEEE*

***Abstract*—The Witsenhausens counterexample is a difficult nonconvex functional optimization problem which has been outstanding for more than 30 years. Considerable amount of literature has been accumulated, but optimal solutions remain elusive. In this paper, we develop a framework that allows us to gain additional new insights to the properties of a better solution for a benchmark instance. Through our approach, we are able to zero in on a solution that is 13% better than the previously known best solution, and more than 54% better than previous results obtained by other authors. More importantly, we demonstrate that our approach, called Hierarchical Search, can be useful in general optimization problems.**

***Index Terms*—Functional optimization, hierarchical search, Witsenhausen counterexample.**

I. Introduction

THE Witsenhausen counterexample [22] has been outstanding for more than 30 years. In this paper, through numerical means, we are able to gain insights about the relationship between the cost terms. The understanding of the cost terms gives us a framework to optimize our solution. Ultimately, we are able to arrive at a solution that is better than any other solution previously known. The outline of the paper is described as follows.

First, we offer a brief background on the Witsenhausen counterexample and a summary of some past attempts. In Section II, we consider a special class of functions—namely the step functions—to represent the possible solutions to the counterexample.[1] We reformulate the cost terms based on the step function framework, which results in a cost objective that is much easier to evaluate. Based on our reformulation of the problem, we present our empirical results and insights, which provide us with the means to zero in on a superior solution in Section III. In Section IV, we attempt to quantify the quality of our solution using results from order statistics. In addition, we compare our solution to solutions obtained by other authors that have yield major improvement, historically. Subsequently, we summarize the knowledge gained in the solution process which applies to other problems that involve searching optimal solution in a large space. We shall conclude in Section VI.

A. The Witsenhausen Counterexample

Team decision problems [9] arise in many engineering and economic systems that typically involve a number of players who are decentralized via some communication channels. Each player chooses actions based on the information received from other players so as to minimize a common cost objective. The important feature that characterizes the difficulty of this problem is the information structure shared among all players. With *classical information structure*, later player has access to accurate information and histories about the observations and actions of earlier players. In the linear quadratic Gaussian (LQG) team problem, it is well-known that classical information pattern enables linear optimal solutions, i.e., the actions taken by each player are an affine function of the available information. Meanwhile, a team decision problem donning *nonclassical information structure*, even if the problem is of the LQG type, has stood as a challenge to researchers for many years.[2] By nonclassical information structure, we mean that any information received by any earlier players is not available to any later players. An outstanding case of this kind is the *Witsenhausens counterexample* [22], which involves only two players.

The common cost objective for the Witsenhausen counterexample is

$$\min_{\gamma_1(\bullet),\,\gamma_2(\bullet)} J = E[k^2u_1^2 + (x + u_1 - u_2)^2] \tag{1}$$

where

x $\sim N(0, \sigma^2)$;
k constant;
u_i $= \gamma_i(y_i)$, $(i = 1, 2)$ action functions chosen by player i.

The information functions y_1 and y_2 are given by

$$y_1 = x \qquad y_2 = x + u_1 + \nu \tag{2}$$

where $\nu \sim N(0, 1^2)$ is the noise corruption to the information received by the second player. We can see from the above that it could be costly for player one to act $(k^2u_1^2)$ despite the perfect information at hand $(y_1 = x)$. On the other hand, player two

Manuscript received September 1, 1998; revised December 1, 1999 and June 1, 2000. Recommended by Associate Editor D. Yao. This work was supported in part by NSF Grant EEC-9527422, Army Contracts DAAL03-92-G-0115 and DAAH04-0148, Air Force Grant F49620-98-1-0387, and the Office of Naval Research Contract N00014-98-1-0720.

The authors are with the Division of Engineering and Applied Sciences Harvard University, Cambridge, MA 02138 (e-mail: {jtlee; ho}@hrl.harvard.edu; twel@post.harvard.edu).

Publisher Item Identifier S 0018-9286(01)02562-4.

[1] A step function is just a piecewise constant function.

[2] Ho and Chu [9] have shown that *partially nested* information structure as a sufficient condition for the optimality of linear solution.

who is able to act without cost considerations cannot compensate player one's action because of the corrupted information of $x + u_1$ by noise ν. Despite the LQG nature, this nonclassical information structure turns the problem into a nonconvex optimization problem in the functional space [i.e., $\gamma_1(\bullet)$, $\gamma_2(\bullet)$] [see (5)]. The optimal solution has been proven not to be of the affine type in Witsenhausens original paper, whereas existence of and conditions for optimum have been determined [22]. Short of the linearity results, this counterexample has invited many attempts in the past three decades, yet optimal solution remains elusive. In Ho and Chang [8], a discrete version of the problem is introduced which admits convex optimization over a severely complicated set of constrained coefficients. This discrete formulation has been later proven NP-complete by Papadimitriou and Tsitsiklis [19].[3]

Prohibitive as it seems, improvements from the linear solution have been reported following the *signaling* concept proposed by Witsenhausen [22]. We shall explain the signaling idea here. First, let us rewrite (1) according to the transformation $f(x) = x + \gamma_1(x)$, $g(y) = \gamma_2(y)$,

$$\min_{f(\bullet),\, g(\bullet)} J(f, g) = E[k^2(x - f(x))^2 + (f(x) - g(f(x) + \nu))^2]. \tag{3}$$

For given $f(x)$ satisfying $E[f(x)] = 0$ and $\mathrm{Var}[f(x)] \leq 4\sigma^2$ (cf. [22, Lemma 1]),[4] Witsenhausen has shown that the *optimal* choice of $g(\bullet)$ is given by

$$g^*(y|f) = \frac{E[f(x)\phi(y - f(x))]}{E[\phi(y - f(x)]} \tag{4}$$

where $\phi(\bullet)$ is the standard Gaussian density. The corresponding objective value becomes

$$J(f) = k^2 E[(x - f(x))^2] + 1 - I(D_f) \tag{5}$$

where

$$I(D_f) = \int \left(\frac{d}{dy} D_f(y)\right)^2 \frac{dy}{D_f(y)} \tag{6}$$

is called the *Fisher information* of the random variable $y = f(x) + \nu$ with density $D_f(y)$. The problem is then converted to minimizing over a single function, namely $f(x)$.[5] We can also see from (5) the nonconvex nature of the cost function because of the convexity of Fisher information [4].[6] Now, in order to help player two distinguish the action of player one from noise corruption, player one could stratify $f(x)$ based on the characteristics of x and ν (essentially their variances) so that player two would be less likely to be mistaken. Effectively, this enhances the information received by player two as reflected in the Fisher information term. This is, of course, not to the best interest of player one because of the incurred expense. However, if k is sufficiently small, then the signaling idea is potentially attractive.[7] Specifically, for $k = 1/\sigma$ and k is small, Witsenhausen has suggested $f(x) = f_w(x)$ to be the signed constant function

$$f_W(x) = \sigma \mathrm{sgn}(x) \qquad g_W(y) = \sigma \tanh(\sigma y) \tag{7}$$

and $g_w(y) = g^*(y|f_w)$ is determined according to (4) above. With this choice of f, player two's information will be either a large positive or large negative number most of the time ($\sigma + \nu$ and $-\sigma + \nu$, respectively). The cost shouldered by player one is given by

$$2k^2\sigma^2 \left(1 - E\left[\left|\frac{x}{\sigma}\right|\right]\right) = 2k^2\sigma^2 \left(1 - \sqrt{\frac{2}{\pi}}\right)$$

and the overall cost objective value is bounded by

$$J(f_W, g_W) \leq 2k^2\sigma^2 \left(1 - \sqrt{\frac{2}{\pi}}\right) + \frac{\sqrt{2\pi}}{k^2}\, \phi\left(\frac{1}{k}\right).$$

For $k \to 0$ and $\sigma = 1/k$, the above bound approaches to only player one's cost which has a value of $J_w = 0.404\,230\,878$, as reported by Witsenhausen [22]. On the other hand, if $f(x)$ is chosen to be affine in x, then the best solution is given by

$$f^*_{affine} = \lambda^* x \qquad g^*(y|f_{affine}) = \lambda^* x$$

where $\lambda^* = (1 \pm \sqrt{1 - 2k^2})/2$. The overall cost objective is $J^*_{affine} \to 1.0$ as $k \to 0$, clearly dominated by J_W. Recently, Baglietto *et al.* [1] reported the ranges and combinations of k and σ such that best affine solutions are dominated by the signaling solutions proposed by Witsenhausen [see (7)]. Essentially, the value of k falls into small values. These problem instances constitute the most challenging and difficult collection. However, more importantly, this counterexample demonstrates that linear solution is not always optimal in LQG problems. It depends on the inherent information structure serving the players.

B. Past Attempts

Since Witsenhausens publication in [22], the signaling concept has been refined via three major avenues: 1) information theory; 2) function approximation; and 3) sampling and search technique. As it will be shown in later sections, this paper furthers the major improvement based on 3), but we shall briefly report efforts in 1) and 2) as follows.

Taking an information theoretic viewpoint, Banal and Basar [2] have optimized the signaling level in $f_w(x)$ from σ to $\sigma\sqrt{2/\pi}$, and the asymptotic best performance is improved from $J_W = 0.404\,230\,878$ to $J^*_{BB} = 0.363$. Another interesting interpretation of signaling using information theory is reported in Mitter and Sahai [18]. They have shown that the solution of the signaling type is infinitely better than the linear solution in the limiting case when $k \to 0$ and $\sigma = 1/k$. As for the function approximation approach, Baglietto *et al.* [1] have used an artificial neural network representation to model $f(x)$. They

[3] In general, the search space in a decision problem can be horrendously large: $|U|^{|Y|}$ where $|U|$ is cardinality of the action/decision and $|Y|$ is the cardinality of the information space. Without exploiting any useful structure, such problems can easily fall into the NP-complete class.

[4] Witsenhausen [22] has shown that the optimal $f(x)$ has to exhibit these properties.

[5] The notation J is slightly abused here, from (3)–(5), since g is completely defined by f as in (5).

[6] The Fisher information $I(D_f)$ is convex [4]. Therefore, $1 - I(D_f)$ is concave. The sum of convex and concave functionals may not be convex.

[7] Later on, we will take advantage of this idea during the construction of our initial solutions.

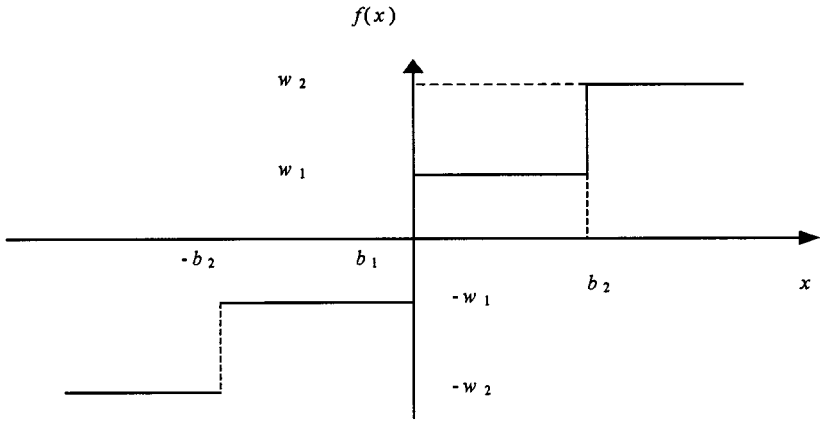

Fig. 1. A 2-step function parametrized by b_1, b_2, w_1, w_2.

have trained a neural network by means of stochastic approximation technique, and obtained results that are comparable to those of Bansal and Basar [2]. More interestingly, the neural network results shown by Baglietto *et al.* demonstrate that the optimal $f^*(x)$ may not be strictly piecewise constant but slightly sloped. This is an important finding, but, nevertheless, it has not been mentioned in [2] or any other work formerly reported! Our results in this paper will show that a $f(x)$ that has a slightly sloped piece between discontinuities, approximated by a series of small steps, performs better than a $f(x)$ that is strictly piecewise constant one. We shall offer explanations based on the cost inefficiencies faced by player two. Our work is inspired by the major numerical advance reported in Deng and Ho [7].

It should be emphasized again that solving this counterexample amounts to finding the best signaling scheme via nonlinear representation of $f(x)$.[8] Deng and Ho [7] have taken a considerably different approach by examining the distribution of performance in the solution space. More precisely, they have employed for the nonlinear representation of $f(x)$—the step functions—parametrizable by a number of break points and signaling levels, respectively, denoted by b's and w's in this paper. For example, the step function $f_W(x)$ proposed by Witsenhausen is a 1-step function—single breakpoint $b = 0$, and signaling level $w = \sigma$. (In Deng and Ho [7], $f(x)$s are constructed to be odd functions, i.e., $f(x) = -f(-x)$. Therefore, the number of steps is counted only in the positive domain of x.) A 2-step function will have two breakpoints b_1 and b_2 in the positive x-axis and two signaling levels w_1 and w_2 in the positive f-axis (see Fig. 1).[9] Notice that step functions allow us to approximate any bounded piecewise continuous function over a compact interval to any degree of predefined accuracy, given that a large enough number of steps is used (whether this is meant for signaling or not, cf. Section I-A. Later in Section III, we will show that it is only necessary to investigate solutions that are finitely supported. This fact together with the continuity property of J (with respect to the parameters) ensure that the step function representation will not miss the global optimum.

[8]We shall elaborate on this in Section III-C.

[9]Due to the symmetry about the origin, the first breakpoint is always zero, i.e., $b_1 = 0$.

In the class of step functions, Deng and Ho [7] have studied the subclasses of 1-, 2-, 3-, 4-, 5-, and 10-step functions. For each of the subclasses, a population of candidate solutions is uniformly generated and put to simulation so as to tally the underlying *performance distribution function* (pdf). The technique of *ordinal optimization* (OO) is employed. It is a method which can locate better solutions with high probability based on rough estimates (e.g., by running shorter Monte Carlo simulations) of the cost function over an array of possible design alternatives (e.g., 1000 different $f(\bullet)$ functions). In other words, OO allows them to evaluate the goodness of a population of designs approximately but quantifiably. (For more information on the concepts, theoretical results and applications of OO, please refer to [7], [10], [15] and [24].) In Deng and Ho's investigation, they have discovered that the pdf associated with the subclass of 2-step functions has higher proportion of "good solutions" compared to the pdf associated with all other subclasses. As a result, the 2-step functions are put into a more in-depth investigation, and eventually they have found a signaling scheme that outperforms all previously reported results by a great margin

$$f_{DH}(x) = \begin{cases} -9.048 & x < -6.41 \\ -3.168 & -6.41 \le x < 0 \\ 3.168 & 0 \le x < 6.41 \\ 9.048 & x \ge 6.41. \end{cases} \tag{8}$$

In our notations, the breakpoints are $b_1 = 0$, $b_2 = 6.41$ and the signaling levels are $w_1 = 3.168$, $w_2 = 9.048$ (see Fig. 1). The corresponding cost objective is measured at $J_{DH} = 0.1901$![10] To understand such a great margin of improvement, notice that the signaling scheme in (8) allows player one to send four messages, i.e., more positive (w_2), less positive (w_1), less negative $(-w_1)$ and more negative $(-w_2)$. Hence, there is a reduction in the magnitude of $(x - f_{DH}(x))$. Meanwhile, the signaling levels are placed sufficiently far apart so that player two can still distinguish player one's messages with small errors, i.e., large Fisher information.

[10]The problem instance considered is $k = 0.2$ and $\sigma = 1/k = 5$. This is also the benchmark case for the Witsenhausen counterexample, and a case mainly focused in this paper. The value $J_{DH} = 0.1901$ is obtained using Monte Carlo simulation method. The associated standard deviation is 0.01.

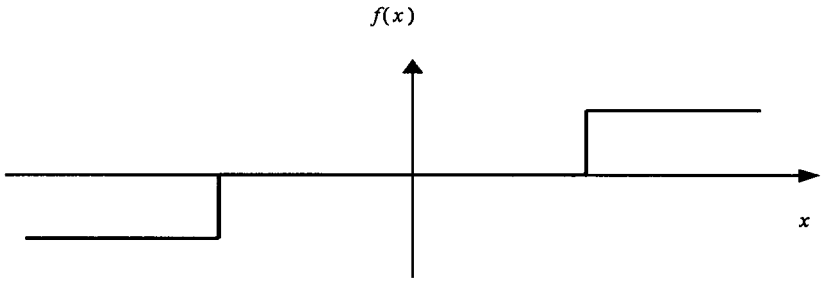

Fig. 2. A 1.5-step function.

In this paper, we examine this signaling scheme in a general framework—we consider the class of n-step functions for representing $f(x)$. By focusing on step functions for $f(x)$, we derive analytical expressions for $g^*(y|f)$ and the cost functional J using only the parameters of breakpoints b's and signaling levels w's. This will be shown in Section II. In Section III, through a benchmark instance of the Witsenhausen counterexample, we analyze the effects of the breakpoints and signaling levels to the stage one and stage two costs.[11] This is done by means of narrowing down the domain and range of $f(x)$. We determine what may constitute the best number of steps and also the parameters of $f(x)$ for the benchmark instance. In addition, by means of the step function formulation, we examine more closely the cost terms and arrive at a surprising and new finding—step functions that resemble piecewise linear functions outperform the purely leveled step functions. More importantly, the approach undertaken in this paper can serve as a model for determining optimal solutions to other problem instances of the Witsenhausen problem. Thus, results of this paper can be viewed as a challenge to additional understanding and meaningful improvement of the solution, continuing the tradition established for this problem for the past 30+ years.

In Section IV, we compare our solution to the ones which have yield major improvement, historically, on the Witsenhausen problem. To contrast, for the benchmark instance, our policy is 13% better than the previously known best. To further substantiate the quality of our solution, we also justify our solution from the sampling point of view. We employ the theory of order statistics that provides us the probabilistic assurance, as shown in Section IV-A.

Last, while our solution is significantly better than any others, it is more instructive to summarize our solution process. In particular, we suggest a hierarchical approach for general functional optimization problems. The search process for better solutions comprises of three mutually supportive components, namely, *goal softening*, *specializing* and *learning*. The work in this paper is an example of such a process and we shall elaborate in Section V.

II. Problem Formulation

The results reported in Deng and Ho [7] suggest that step functions can help reducing the overall cost objective by a great margin. We decide to pursue the step functions further in this paper. Notice that the class of step functions can approximate any target function satisfying certain regularity conditions, e.g., piecewise continuous, to any degree of accuracy when sufficient number of breakpoints b's and signaling levels w's are given. Furthermore, it can be shown that the objective function J is continuous with respect to the parameters b's and w's (see Appendix I). This assures us that if a step function closely approximates the global optimum, the objective value of this candidate solution is going to be near the objective value associated with the global optimum. Therefore, by focusing on the class of step functions, we will not leave out the global optimum during our search but up to numerical accuracy.

Before we start to formulate the problem using step functions, the following are the assumptions on the optimal $f^*(\cdot)$:

1) $f^*(\cdot)$ is symmetric about the origin;
2) $f^*(\cdot)$ is monotone nondecreasing.

Property 1 is based on [22, Lemma 1] that $E[f^*(x)] = 0$. (See Appendix III for more rationale of this assumption.) Property 2 is proven by Witsenhausen in [22, Lemma 7]. These properties help us to narrow down in the solution space of all one-dimensional functions in our search. Unless otherwise stated, we will be considering in this paper only those $f(x)$ which exhibit these properties.

A. "Half" Step Functions

Given that $f(x)$ is an odd function, it is symmetric about the origin and can be completely defined by specifying only in the positive domain. Thus, the number of steps of such an odd function is counted only in the positive domain. In this fashion, a 2-step function $f(x)$ is in fact having four steps in the entire domain of $f(x)$. Notice that it is possible to have an n-step function with the first signaling level to be zero, i.e., $w_1 = 0$. By symmetry, the first step in the positive domain is actually "half" of a step about the origin (see Fig. 2 below). To distinguish this type of step functions, we denote it as "$n.5$-step" function. Fig. 2 contains an example of a 1.5-step function. Note that a 0.5-step function is the same as $f(x) = 0$, for all x. As for indexing, we adopt the convention that the breakpoints are labeled from $i = 1, \ldots, \lfloor n + 0.5 \rfloor$ for an "n-step" or "$n.5$-step" function.

B. Problem Formulation

The version of the Witsenhausen counterexample considered in this paper can be formally stated as follows:

$$\min_{f \in \Theta} J(f) = E[k^2(x - f(x))^2 + (f(x) - g^*(f(x) + \nu))^2]$$

[11] We use the terms *player one's cost* and *stage one cost* and the terms *player two's cost* and *stage two cost* interchangeably.

where

x $\sim N(0, \sigma^2)$;
k $= 1/\sigma$;
ν $\sim N(0, 1)$;
Θ set of all nondecreasing step functions.

Notice that an instance is completely specified by the value of σ when $k = 1/\sigma$. The optimal response of player two, $g^*(f+\nu)$, is given by Witsenhausen as shown in (4). Given that $f(x)$ is a step function, we can derive the closed-form expression for $g^*(f+\nu)$. The cost terms in $J(f)$ based on step function f and corresponding g^* will be shown subsequently.

1) Expression for $g^*(y)$: An n-step function $f(x)$ is parametrized by a $(2n+1)$-tuple $(n; b_1, \ldots, b_n; w_1, \ldots, w_n)$, and for it is nondecreasing then we have $0 = b_1 \le b_2 \le \cdots \le b_n < \infty$ and $0 \le w_1 \le w_2 \le \cdots \le w_n < \infty$. The function $f(x)$ can be written as

$$f(x) = -w_n + \sum_{i=1}^{n-1}(w_{i+1} - w_i)T(x + b_{i+1}) + 2w_1 I(x) + \sum_{i=1}^{n-1}(w_{i+1} - w_i)T(x - b_{i+1})$$

where $T(\bullet)$ is the standard unit-step function. Notice that the image of $f(x)$ has only $2n$ point masses at $-w_n, \ldots, -w_1, w_1, \ldots, w_n$. The probabilities at these point masses are, respectively, $p_n, \ldots, p_1, p_1, \ldots, p_n$, where

$$p_i = \int_{b_i}^{b_{i+1}} \phi(s; 0, \sigma_0^2)\, ds = \frac{1}{2}\left(\operatorname{erf}\left(\frac{b_{i+1}}{\sqrt{2\sigma^2}}\right) - \operatorname{erf}\left(\frac{b_i}{\sqrt{2\sigma^2}}\right)\right) \quad \text{for } i = 1, \ldots, n$$

and $b_{n+1} \equiv \infty$, $\phi(\bullet; 0, \sigma^2)$ is a Gaussian density with mean zero and variance σ^2, and $\operatorname{erf}(\bullet)$ is the standard error function. Using the optimal form provided by Witsenhausen [22] and let $y = f + \nu$, we have

$$g(y) = g^*(y|f) = \frac{E_x[f(x)\phi(y - f(x))]}{E_x[\phi(y - f(x))]}. \tag{9}$$

Expanding the numerator of $g(y)$, we get

$$-\sum_{i=1}^{n} p_i w_i \phi(y + w_i) + \sum_{i=1}^{n} p_i w_i \phi(y - w_i)$$

and, similarly, the denominator

$$\sum_{i=1}^{n} p_i \phi(y + w_i) + \sum_{i=1}^{n} p_i \phi(y - w_i).$$

Making the following substitution

$$\phi(y - w_j) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2} + w_j y - \frac{w_j^2}{2}\right)$$

we have

$$g(y) = \frac{\sum_{i=1}^{n} p_i e^{-w_i^2/2}[e^{w_i y} - e^{-w_i y}]w_i}{\sum_{i=1}^{n} p_i e^{-w_i^2/2}[e^{w_i y} + e^{-w_i y}]}.$$

By writing $[e^{w_i y} - e^{-w_i y}] = 2\sinh(w_i y)$ and $[e^{w_i y} + e^{-w_i y}] = 2\cosh(w_i y)$, we get

$$g(y) = \frac{\sum_{i=1}^{n} p_i \phi(w_i) \sinh(w_i y) w_i}{\sum_{i=1}^{n} p_i \phi(w_i) \cosh(w_i y)}.$$

Now that we have a closed-from expression for $g(y)$, no longer do we need to evaluate $g(y)$ by simulation [due to the expectations in (4)], which could be costly.

2) Expression for the Cost Term $J(f)$: Since $f(x)$ is a step function, it is easy to compute the stage one expected cost, i.e.,

$$k^2 E[(x - f(x))^2] = 1 - \frac{2}{\sqrt{2\pi\sigma^2}} \sum_{i=1}^{n} w_i \left[\exp\left(\frac{b_i^2}{2\sigma^2}\right) - \exp\left(\frac{b_{i+1}^2}{2\sigma^2}\right)\right] + 2\sum_{i=1}^{n} p_i k^2 w_i^2. \tag{10}$$

Now, let us bring in the expression for the probability density function of y, i.e., $D_f(y)$. Since $y = f(x) + \nu$, then $D_f(y)$ is the convolution of the point masses of $f(x)$ and the density of ν, which gives [recall that $\nu \sim N(0, 1)$]

$$D_f(y) = 2\sqrt{2\pi}\phi(y) \sum_{i=1}^{n} p_i \phi(w_i) \cosh(w_i y). \tag{11}$$

Unfortunately, even equipped with the knowledge of $D_f(y)$, there is still no closed-form expression for the Fisher information of y. Therefore, the second stage expected cost $E[(f(x) - g(f(x) + W))^2]$ has to be computed by inserting Equation (11) into (6) which is then numerically integrated. Combining (10) and the numerical result of the Fisher information of y, we have a way to evaluate the objective value

$$J(f) = \left\{1 - \frac{2}{\sqrt{2\pi\sigma^2}} \sum_{i=1}^{n} w_i \left[\exp\left(\frac{b_i^2}{2\sigma^2}\right) - \exp\left(\frac{b_{i+1}^2}{2\sigma^2}\right)\right] + 2\sum_{i=1}^{n} p_i k^2 w_i^2\right\} + \{1 - I(D_f)\} \tag{12}$$

for any given $f(x)$ being a step function.

Another property of the cost function J is that it is continuous with respect to the parameters of the step function, i.e., $(b_1, \ldots, b_n; w_1, \ldots, w_n)$, which is shown in Appendix I. We

TABLE I
COST FOR VARIOUS BREAKPOINTS

	Stage 1 Cost	Stage 2 Cost	Total Cost
$b_2 = 4.41$	0.203898117969	0.056907689938	0.260805807906
$b_2 = 6.41$ (DH)	0.139488401953	0.053072896985	0.192561298938
$b_2 = 8.41$	0.200258857650	0.045308479077	0.245567336727

TABLE II
COST OF BEST $f(x)$ FOUND DURING THE INITIAL SEARCH

n-step	Stage 1 Cost	Stage 2 Cost	Total Cost
0.5-step	1.0	0.0	1.0
1-step	0.363530423422	0.001299586561	0.364830009984
1.5-step	0.198843219959	0.013408849464	0.212252069424
2-step	0.151787982368	0.025679977982	0.177467960350
2.5-step	0.142795089815	0.029190918087	0.171986007902
3-step	0.140892490172	0.030763466092	0.171655956264
3.5-step	0.142122494792	0.029272149650	0.171394644442
4-step	0.140834875577	0.030792110476	0.171626986053
4.5-step	0.142073541452	0.029453477715	0.171527019167

TABLE III
COST OF BEST $f(x)$ FOUND IN THE BOUNDED DOMAIN AND RANGE

n-step	Stage 1 Cost	Stage 2 Cost	Total Cost
0.5-step	1.0	0.0	1.0
1-step (B&B [2])	0.363380227653	0.001634601425	0.365014829078
1.5-step	0.198843219959	0.013408849464	0.212252069424
2-step	0.151787982368	0.025679977982	0.177467960350
2.5-step	0.142795089815	0.029190918087	0.171986007902
3-step	0.140892490172	0.030763466092	0.171655956264
3.5-step	0.142122494792	0.029272149650	0.171394644442
4-step	0.140834875577	0.030792110476	0.171626986053
4.5-step	0.142073541452	0.029453477715	0.171527019167
5-step	0.110208380868	0.087097601191	0.197305982060
5.5-ste-	0.090133335704	0.167843992337	0.257977328041
5-step	0.073633338722	0.280781920757	0.354415259478
6.5-step	0.061633339960	0.399569880799	0.461203220759

shall utilize this property in our search of optimal $f^*(x)$ in the later sections.

C. A Note on Computational Savings

The formulation in the previous section allows us to numerically evaluate the cost objective instead of estimating it by simulation. Typically, it may take up to 3–4 h using Monte Carlo simulation to get a single cost estimate with standard deviation of 0.0001 for one particular $f(x)$.[12] Now, we can compute the cost objective of 5000 2-step functions in about 40 s with a choice of integration step size of 0.05 (the accuracy in the numerical integration is within an order of 10^{-4}). In other words, with the new computational scheme, we only need 1/1 350 000 the amount of time to compute the performance index associated with one $f(x)$ to the same degree of accuracy. This allows us to evaluate a lot more designs in a shorter period of time.

To contrast the computational scheme using ordinal optimization in Deng and Ho [7], they use short Monte Carlo simulations to obtain quick but "rough" estimates of the cost objective. On the other hand, we use approximate cost objective associated with each $f(x)$.[13] The commonality between the two computational schemes is that both are fast. The computational speed-up is what allows both Deng and Ho [7] and us to search directly in the space of step functions.

III. EMPIRICAL RESULTS AND INSIGHTS

Based on the framework developed above, the problem becomes identifying what signals (b_i, w_i), $i = 1, \ldots, n$, to be sent so as to minimize player one's cost and to maximize the Fisher information on $f + \nu$. During our investigation, we have focused our attention on the benchmark instance where $\sigma = 5$ and $k = 0.2$. We implemented a form of the scatter search technique (see Appendix II) to achieve local improvements from a given candidate solution. More importantly, as in many search methods, knowledge pertaining to the specific problem should be exploited in the construction of (initial) solutions. Our purpose here is to identify some prominent features using empirical data in our study. These features include

1) location of breakpoints b_i's with respect to the signaling levels w_i's;
2) coverage of the domain and range of $f(x)$.

We will be utilizing these features to construct initial solutions, which are then iterated by the scatter search technique. Subsequently, our analysis will be in two folds: we first determine how many steps will strike the best balance between the player one and player two costs, then we examine the cost inefficiencies in each stage for further improvements.

A. Placement of Breakpoints

Our first observation is that the location of breakpoints should be approximately the average of the two adjacent signaling levels. This means that $b_i \approx (w_{i-1} + w_i)/2$, for $i \leq 2$ (recall that $b_1 = 0$ due to symmetry). To see this, notice that player one's cost is $k^2 E[(x - f(x))^2]$. Suppose player one had focused solely on the stage one cost term, it is then straight forward to show that the *optimal* breakpoint b_i for player one is exactly at $(w_{i-1} + w_i)/2$. However, in view of the cost shouldered by player two, the breakpoints are in general not exactly the average of the adjacent signaling levels but this may serve well as the initial solution for search purpose.

To illustrate this further, we take the best 2-step function report in Deng and Ho [7], and we vary the breakpoint from $b_2 = 6.41$ to 4.41 and to 8.41, while keeping the adjacent signaling levels $w_1 = 3.168$ and $w_2 = 9.048$ unchanged.[14] We compare the cost objectives in these three cases in Table I.[15] There is a significant increase in the stage one cost when the breakpoint is moved to the left or to the right of the breakpoint $b_2 = 6.41$. This can be easily seen in the increased magnitude of $(x - f(x))$, as in Fig. 3. Notice that the changes in stage two cost are less pronounced compared to that of stage one. Therefore, it is reasonable to embark the search by locating breakpoints around the average of the two adjacent signaling levels. What remains to be determined is the placement of signaling levels.

[12]This is done on a SUN Sparc 20 workstation.

[13]Accuracy of the scheme depends on the algorithm used to numerically evaluate the integral of the Fisher information term as well as the number of significant digits in the computer.

[14]Notice that $b_2 = 6.41$ is greater that the average of $w_1 = 3.168$ and $w_2 = 9.048$.

[15]Entries in the row $b_2 = 6.41$ (DH) are calculated by numerical integration as derived in Section II. They differ from the simulation results reported in Deng and Ho [7] but only for a small margin.

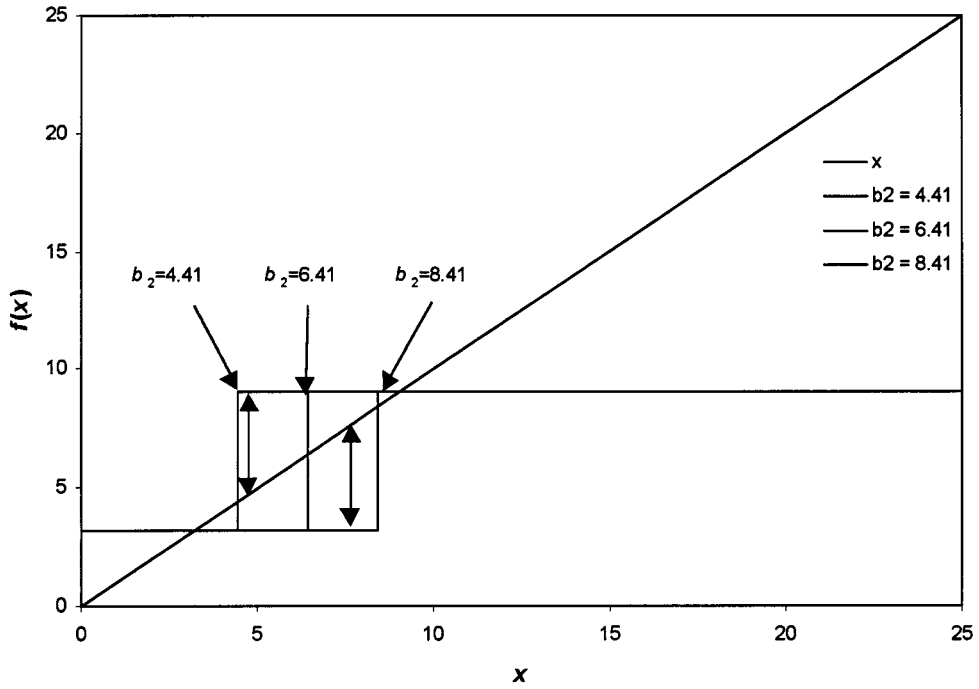

Fig. 3. $f(x)$ with breakpoints $b_2 = 4.41$, 6.41 (DH) and 8.41 (see also Table I).

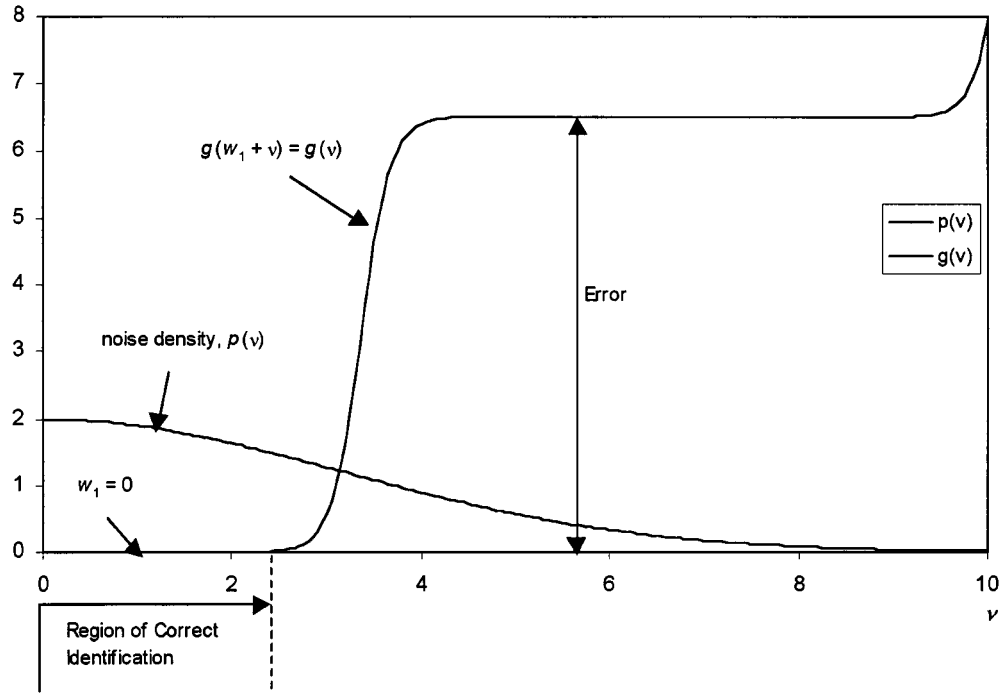

Fig. 4. Origin of mis-identification error for signal $w_1 = 0$ [$f(x)$ is a 3.5-step function here].

B. Coverage of Domain and Range of $f(x)$

Table II summarizes some preliminary results in our initial study. We begin by successively adding from $f(x) = 0$ (i.e., a 0.5-step function) signaling levels with equal separations.[16] (For the benchmark instance $\sigma = 5$, the separation is about 3.0.) Then, for each of these step functions, we improve locally using scatter search technique.

One observation from Table II is that as the number of steps increases, the total cost starts to converge around 0.1715. In view of the probability distribution of x, i.e., $N(0, \sigma^2)$, when defining one more step for $f(x)$ beyond the domain $[-5\sigma, 5\sigma]$ (i.e., $[-25, 25]$ in this case), the effects of this extra step to the probability masses of w_i's and the overall cost objective are by and large very insignificant. Therefore, the total costs become

[16]Equal separation between adjacent steps is only imposed during the construct of an initial design. There is a large enough separation between adjacent signals thus to insure the Fisher information terms is maximized. Hence, a lower stage two cost will result.

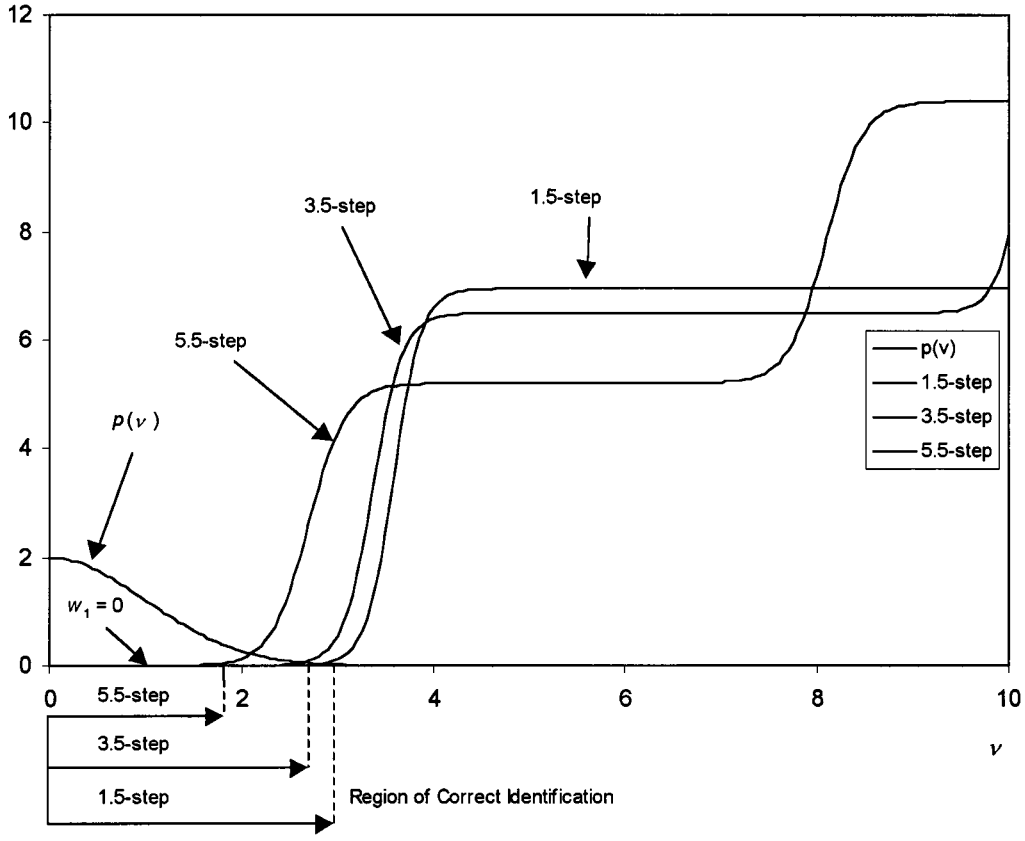

Fig. 5. Comparing the regions of correct identification for 1.5-, 3.5-, and 5.5-step functions (figure not drawn to scale).

numerically indistinguishable. In addition, Witsenhausen has shown that the optimal $f^*(x)$ must satisfy $E[(f^*(x))^2] \leq 4\sigma^2$ (cf. [22, Lemma 1]). As a result, we are going to study step functions in a bounded domain and range in the next section.

C. Step Functions in the Bounded Domain and Range

In this section, we investigate different $f(x)$s by adding more signaling levels into the range of $[-5\sigma, 5\sigma]$. We also pay special attention to the trade-offs between the stage one and stage two cost terms when more steps are inserted. It is then possible for us to narrow down the appropriate number of steps that strikes the best balance between two stages. Table III summarizes our finding. Each row in the table is constructed from an initial solution but local improvements are made by means of scatter search technique.

A direct consequence of inserting more signals is that they are "crowding up" the bounded range as the number of steps increases. On one hand, this reduces the stage one cost as shown in the second column in Table III. Since the breakpoints are around the average of adjacent signaling levels, when more steps are inserted, $f(x)$ as a step function will zig-zag more closely around x, i.e., $f(x)$ better approximates x. Hence, the first stage cost $k^2E[(x - f(x))^2]$ is decreased. On the other hand, when more signals are sent by player one in the bounded range, player two will have more difficulties in discerning different signals because of the decreased separation and the underlying observation noise ν. In fact, this more refined signaling scheme increases the chance that a signal is being misidentified which will in turn cost player two.

To illustrate further, we can examine the stage two cost again in (13) [p_i is the probability of $f(x) = w_i$].

$$E[f(x) - g(f(x)+\nu)]^2 = 2\sum_{i=1}^{n} p_i E_v[w_i - g(w_i+\nu)]^2. \quad (13)$$

Consider the case when $i = 1$ and $w_1 = 0$ (i.e., the first signaling level of a $n.5$-step function), the misidentification error in stage two is given by

$$\begin{aligned} p_1 E_v[(w_1 - g(w_1+\nu))^2] &= p_1 E_v[g^2(\nu)] \\ &= p_1 \int_{-\infty}^{\infty} p(\nu)[g(\nu)]^2 \, d\nu \end{aligned}$$

where $\nu \sim N(0, 1)$ is the noise in the channel. Fig. 4 shows the region of correct identification and origin of error for the case of the best 3.5-step in Table III. The stage two cost is incurred in the area where $g(\nu)$ is not zero, i.e., away from the region of the correct identification. After squaring the error terms, they are then weighted by the density function of ν, $p(\nu)$. If there are fewer number of steps, then the region of correct identification is enlarged. On the other hand, more steps would shrink the region. This can be seen in Fig. 5 where the $g(\nu)$ functions for the best 1.5-, 3.5- and 5.5-step functions are plotted. As the number of steps increases, the region of correct identification shrinks,

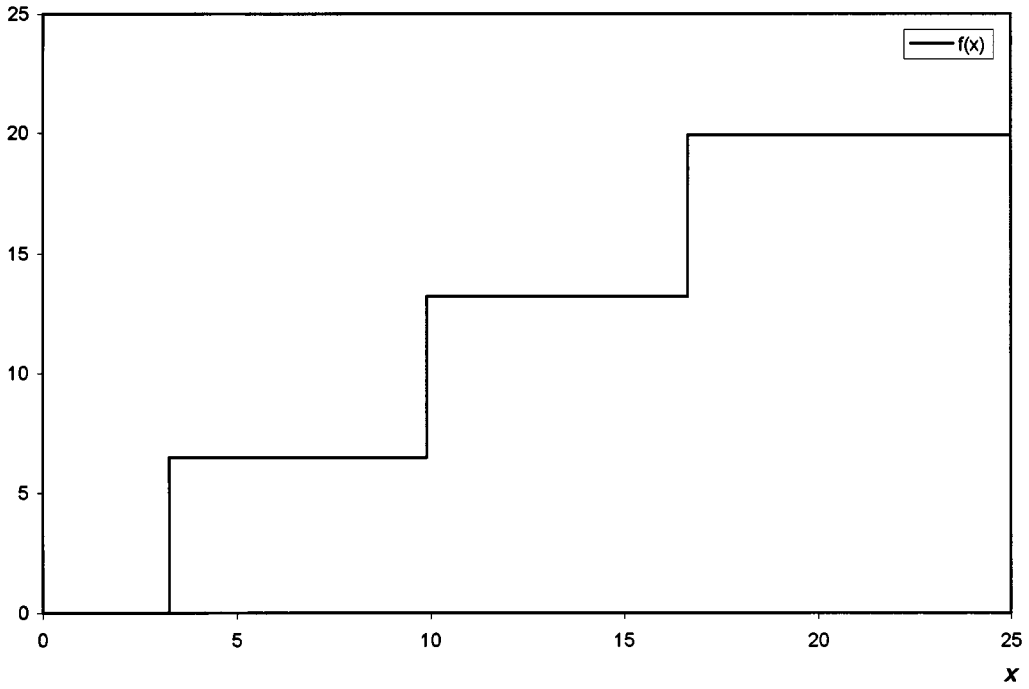

Fig. 6. Best 3.5-step function.

and the error of mis-identifying the signal $w_1 = 0$ increases. The same argument applies to other signaling levels w_i, $i \geq 1$. Overall speaking, the stage two cost increases despite better approximation of x by $f(x)$ in stage two.

Therefore, in order to bring down the stage two cost, it is desirable to create greater separations between the adjacent signaling levels. Fewer number of steps in a bounded range for $f(x)$ will allow greater separations. However, as we have seen earlier, this results in poorer approximation of x and increases the stage one cost. Finding the optimal $f^*(x)$ in the class of step functions thus amounts to finding the optimal number of steps in $f(x)$ and their placements so as to balance the tradeoffs between the first and second stage costs.

From the numerical results in Table III, the 3.5-step seems to strike the best balance between the trade-offs in minimizing stage one and stage two costs. Thus, the signaling scheme allows player one to send seven signals to player two, i.e., large negative (-19.9), medium negative (-13.2), small negative (-6.5), zero, small positive (6.5), medium positive (13.2) and large positive (19.9). Equation (14) below defines the best 3.5-step function that we have found [due to symmetry, $f(x)$ is only shown for $x \geq 0$]. Fig. 6 shows a plot of (14)

$$f(x) = \begin{cases} 0 & 0 \leq x < 3.25 \\ 6.5 & 3.25 \leq x < 9.90 \\ 13.2 & 9.90 \leq x < 16.65 \\ 19.9 & 16.65 \leq x. \end{cases} \tag{14}$$

D. Toward Piecewise Linear Step Functions

Can we still do better? In our earlier attempt to solve the Witsenhausen counterexample, it has been found that if a step function is slightly sloped in each of the steps, the cost objective can be further improved from those which have perfectly leveled steps. As we shall see momentarily, this improvement originates from the quadratic nature of the cost terms in the two stages. Let us pursue the idea of slightly sloped step functions first. In order to make use of the step function formulation developed in Section II, we need to approximate the slightly sloped steps accordingly.[17] We take the best 3.5-step function in the last section [i.e., (14)], and break each of the 3.5 steps into smaller staircase-like segments which "track" each original signal in an increasing fashion. We also simplify the insertions to be equally spaced, both horizontally and vertically, so that the segments appear to increase "linearly" (see Fig. 7). The insertions are at a distance δ from the signaling level and from each successive segment. The value of δ is optimized at each signaling level over the number of segments added. Our observations are shown in Table IV; the more the number of segments in each step, the lower the total cost measured. It turns out that more improvements can be made by carefully adding segments to a perfectly leveled step. When the number of insertion increases, the collection of segments gradually approaches a slightly sloped line segment.

To understand these improvements, we need to review the nature of the cost components $E[(x - f(x))^2]$ and $E[(f(x) - g^*(f(x) + \nu))^2]$; both of them are second order polynomials in x, f, and g^*. For a quadratic function, the further away it is from its extremum, the higher the rate of change of the function in magnitude. For the 3.5-step function shown in (13), the stage one cost is a lot higher than the stage two cost. By breaking down each signaling level into smaller segments of increasing values,

[17]Due to the continuity property of J with respect to the breakpoints and the signaling values, there is no loss of generality to approximate the slightly slopped steps by a step function.

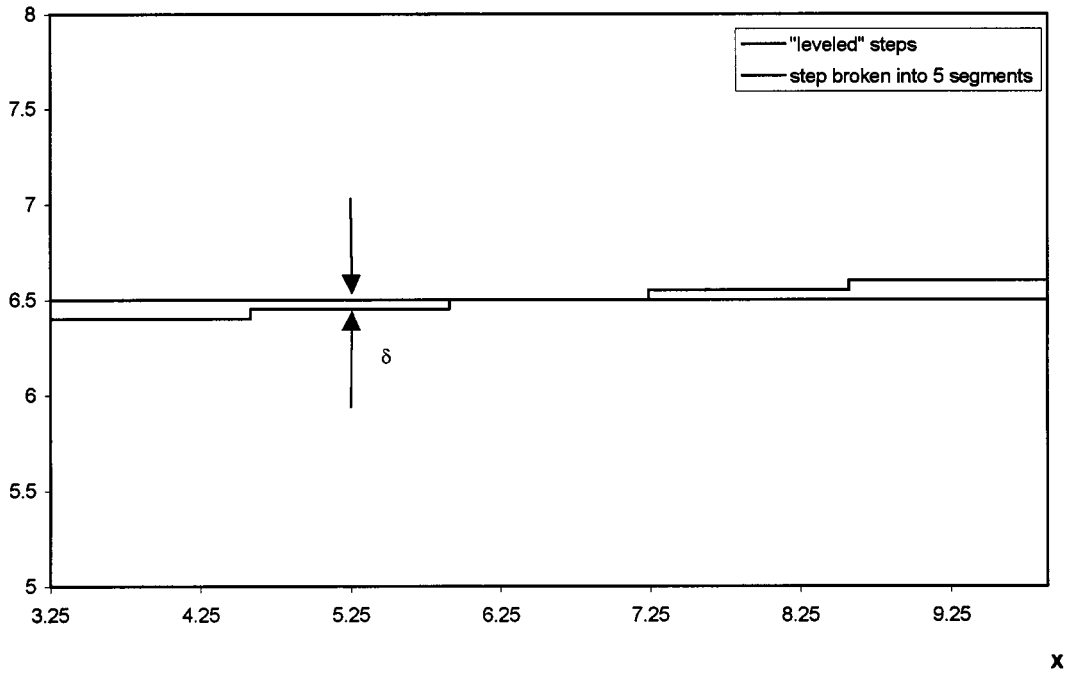

Fig. 7. Breaking up a "leveled" steps into five segments.

TABLE IV
REFINEMENT OVER 3.5-STEPS BY INSERTING STAIRCASE-LIKE SEGMENTS

	Stage 1 Cost	Stage 2 Cost	Total Cost
Leveled	0.142122494792	0.029272149650	0.171394644442
2 segments	0.135703004561	0.032581407191	0.168284411752
3 segments	0.134133682395	0.033489499564	0.167623181960
4 segments	0.132521124885	0.034876169908	0.167397294793
5 segments	0.131884081844	0.035429123524	0.167313205368

the stage one cost decreases because this reduces the absolute differences between x and $f(x)$. Since the cost components are quadratic, the reduction in the stage one cost outweighs the increase in the stage two cost. Therefore, we can still edge in more improvements in the overall cost.

One might wonder if the finding above contradicts our earlier insight that too many signals will confuse player two. The answer is no. The addition of these segments is *not* the same as adding more signaling steps because of the segments' proximity. Since player two has a noisy observation, the little segments with similar values will be perceived as one signal over a noisy channel. The small variations between the small segments will be smudged by noise. Given that each cluster of small segments have large separation between them, player two will still be able to distinguish the different signals.

It is also appropriate to contrast our findings here with those reported by Baglietto *et al.* [1] who have used an artificial neural network approach for the Witsenhausen counterexample. More specifically, they have constructed neural networks to serve as the nonlinear approximators for $f(x)$ and $g(f(x)+\nu)$. It is well known that neural network possesses very good function approximation properties, and it is widely used in classification problems. In Baglietto *et al.* [1], they have used stochastic approximation to obtain a neural network for $f(x)$, and the network output is similar to a one-step slightly sloped piecewise linear function as discussed in this section. This neural network allows Baglietto *et al.* [1] to report better results than that of Banal and Basar [2], mainly due to the sloped effect and the reason we have just outlined. [18] However, it seems that the neural network approach cannot be easily modified to arrive at a multistep function proposed in this paper, unless a clever topology can be devised.

IV. COMPARISON TO THE HISTORICAL ATTEMPTS

By means of the step function formulation in the previous sections, we have achieved the following.

1) A fast and accurate computational scheme for the total cost $J(f)$ which eliminates the need of simulation, but requires numerical methods in evaluating the Fishers information $I(D_f)$ has been acheived.
2) The insights on the placements of breakpoints and signaling levels that reduce the respective cost components in stage one and stage two. In particular, breakpoints are located around the value of the average of adjacent signaling levels so as to bring down the stage one cost. A total of 3.5 signals are placed in a bounded range of $f(x)$ so as to allow enough separation for player two to discern the signals in the presence of noise. Furthermore, additional improvement can be made by adding small segments to each signal. This is taking advantage of the quadratic nature of the cost terms.
3) A solution which is at a 13% improvement from the best solution in Deng and Ho [7], which is 47% better than

[18]The authors became aware of some new results from Baglietto *et al.* after this paper was accepted for publication. For more information, refer to [26].

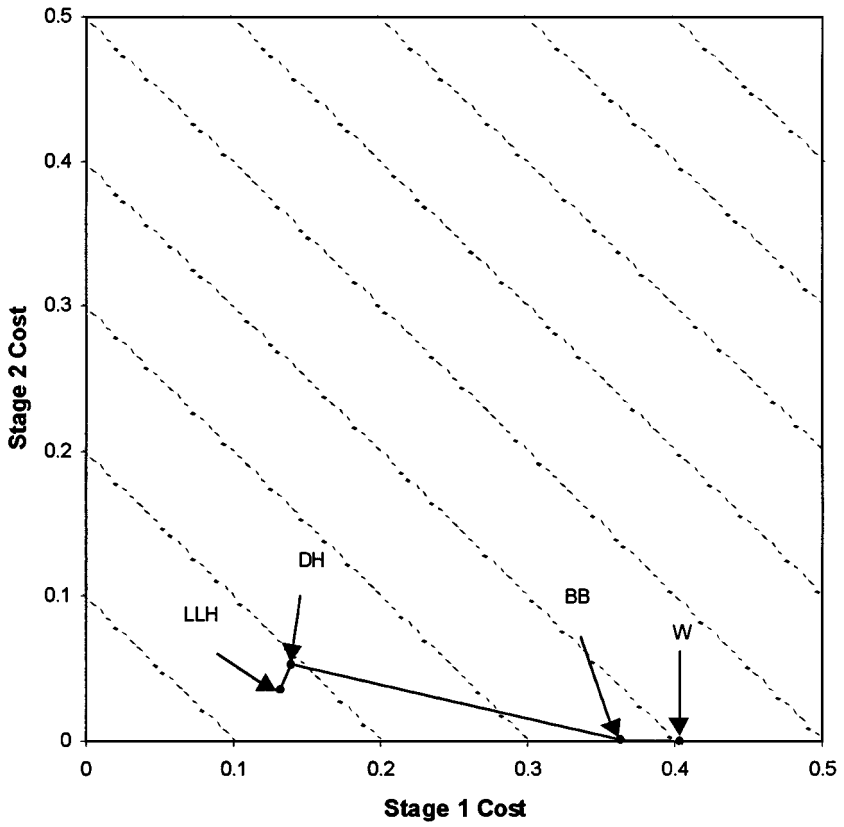

Fig. 8. Historical improvements on the Witsenhausen counterexample (benchmark: $k = 0.2$ and $\sigma = 5$).

the solution found by Banal and Basar [2], i.e., a 3.5-step function with added segments in each step

$$f(x) = \begin{cases} 0.00 & 0.00 \le x < 0.65 \\ 0.05 & 0.65 \le x < 1.95 \\ 0.10 & 1.95 \le x < 3.25 \\ 6.40 & 3.25 \le x < 4.58 \\ 6.45 & 4.58 \le x < 5.91 \\ 6.50 & 5.91 \le x < 7.24 \\ 6.55 & 7.24 \le x < 8.57 \\ 6.60 & 8.57 \le x < 9.90 \\ 13.10 & 9.90 \le x < 11.25 \\ 13.15 & 11.25 \le x < 12.60 \\ 13.20 & 12.60 \le x < 13.95 \\ 13.25 & 13.95 \le x < 15.30 \\ 13.30 & 15.30 \le x < 16.65 \\ 19.90 & 16.65 \le x \end{cases} . \quad (15)$$

4) The corresponding total cost is measured at $J = 0.167\,313\,205\,338$. The added segments further reduce the stage one cost while the stage two cost is only barely affected. In the limit, these added segments will produce a slightly sloped piecewise linear solution for $f(x)$.

It is also useful to contrast the historical attempts on the Witsenhausen counterexample which is shown in Fig. 8 for the benchmark $\sigma = 5$ and $k = 0.2$. The horizontal and vertical axes correspond to the stage one and stage two costs respectively whereas the dashed lines represent the iso-cost contour, i.e., total cost = stage one cost + stage two cost. In the figure, "W" stands for the total cost published by Witsenhausen in his original paper [22]. "BB" is the total cost given by 1-step function found by Banal and Basar [2] which has a signaling level at $\sigma\sqrt{2/\pi}$. "DH" represents the total cost of the best 2-step function reported in Deng and Ho [7]. Last, "LLH" marks the total cost of the best solution found in this paper. These total costs are also compared in Table V.[19]

As shown by the points marked "W," "BB," and "DH" in Fig. 8, we see that historically improvements have been made on reducing the stage one cost but sacrificing the stage two cost. Meanwhile, our work has brought both the stage one *and* stage two costs down, which is possible by identifying 3.5-step provides the best balance between the two stages, and the addition of small segments which edge in more improvements in stage one.

It is obvious that the only possibility to outperform our solution is to improve the stage one or stage two cost, or both. From the insights gained in our investigation, we argue that the stage one cost cannot be improved except in the limiting case by admitting slightly sloped steps, which can only bring forth marginal and numerical improvements. Similarly, the stage two

[19]Same as footnote 13.

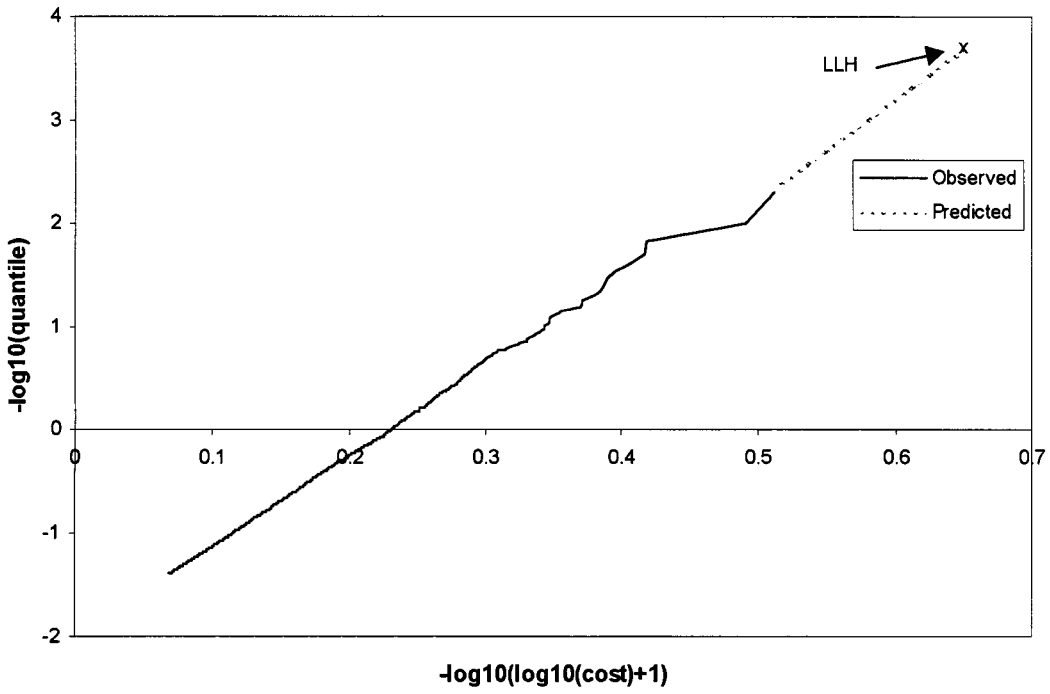

Fig. 9. Regression result based on the top 5000 samples out of 20 000 samples.

cost cannot be improved further without elevating the stage one cost, i.e., there is neither room to "sharpen" the signals by allowing further separations, nor any reduction in the number of signals. Thus, results of this paper can be viewed as a challenge to additional insights and meaningful improvement of the solution.

A. Quality of Solution

Since our attempt in this paper is nonetheless numerical, it will be useful to gauge in some way the quality of our solution. Specifically, we employ results from *the order statistics* literature. We will first introduce briefly some preliminaries, but readers interested in the theory of order statistics may refer to David [6].

First, notice that the best solution shown in (15) is in fact in the category of 14-step functions. Imagine that one had evaluated all possible solutions in the space of 14-step functions, denoted by Θ_{14}, and tallied all these cost values in a histogram. This can be considered as what is called the "performance density function" in Deng and Ho [7]. The histogram serves as the performance density function for the space Θ_{14} when proper normalization is performed.

Second, suppose one samples uniformly a collection of N 14-step functions from the space Θ_{14}. Each of them is indeed a possible solution to the Witsenhausen problem (but probably all being suboptimal). When the performances of these N functions have been evaluated, they are ordered from the smallest to the largest, forming a total of N ordered performances. These ordered performances can be regarded as estimates of all the N order statistics from the performance density function of Θ_{14}. It is known in the literature that these N order statistics divide up the performance density function into $(N+1)$ equal parts, on the average. The smallest order statistic is commonly referred to as the *extreme value* statistics.

Now, we want to answer the following question: What is the chance that the best of these N samples will fall into the top q% of the entire solution space Θ_{14}? In other words, what is the probability that the cost value of the best sampled 14-step function will be lower than the q% population quantile, say, ξ_q? By standard results in order statistics, this probability is given by[20]

$$\text{Prob(performance of the best sampled solution} < \xi_q) = 1 - (1 - q\%)^N.$$

In this study, we have sampled uniformly $N = 20\,000$ different 14-step functions from the space Θ_{14}. The best performance of these 20 000 solutions is measured at 0.203 16. At a confidence level of 0.9999, we solve from the order statistics result above that this solution is among the top 0.046% of the entire population. In other words, 0.203 16 is an estimate for the top 0.046% quantile. Meanwhile, the solution posted in (15) has a cost value measured at 0.167 313 205 338. Therefore, we assert with a confidence level of 0.9999 that the our best solution in (15) is among the top 0.046%, if not better, in the space of 14-step functions Θ_{14}.

From the 20 000 samples above, we can also predict the quality of our best solution in the space of 14-step functions Θ_{14} using regression techniques. Since the 20 000 order statistics on average divide up the performance density function into 20 001 equal parts. In other words, the cost value associated

[20]Results in order statistics require the underlying density to be continuous, which is the case we have here as we show Appendix I the continuity of performances J with respect to the parameters b_i and w_i in the step functions.

with the ith ordered sample on average marks the quantile 100 $(i/20\,001)$% of the performance value in the space Θ_{14}. If we plot the measured cost values associated with the 20 000 ordered samples in the horizontal axis and plot the evenly spaced quantile marks above in the vertical axis, one can think of this as an estimate of the "performance distribution function" of Θ_{14} as in Deng and Ho [7].[21] After scaling both the horizontal and vertical axes, we observe a linear relationship between the costs and the quantiles in the region of the 5000 samples with the smallest cost values.[22] We perform linear regression based on those 5000 data points. The linear regression model predicts that the cost value associated with our best solution 0.167 313 205 338 is among the top 0.000 205% $\pm$ 0.000 004%, where 0.000 004% represents one standard deviation of possible error, in the space of 14-step functions Θ_{14}.

V. Hierarchial Search Method

While what we have gained significant insight about a 30-year old puzzle, it is actually more instructive, from an algorithmic point of view, to highlight the search process conducted throughout this study. In this section, we summarize our strategy and efforts into a more general framework. In searching for the global optimum, we employ what we called a *hierarchical search method*. Specifically, we have adopted a two- stage hierarchy in this problem. Let us explain as follows.

In the first stage of the hierarchy, we survey a higher level question: How many (major) steps are needed? We have tested from each subclass of n-step functions, $n = 0.5, 1, 1.5, \ldots, 6.5$, and discover empirically from data that 3.5-step functions strike the best balance among others. In the second stage of the hierarchy, we analyze what goes into the cost inefficiencies in each stage, and become convinced that the insertion of smaller steps which approximates a slightly sloped segment is indeed advantageous.

Meanwhile, functional optimization remains as a very challenging and central area in the study of decision and control. The most celebrated example is the LQG control problem which admits not only closed form solutions but also a handful of techniques for solving them (e.g., dynamic programming, direct methods, etc., see Bryson and Ho [3]). We set out in this paper with a team decision problem which is LQG in nature but augmented only with the nonclassical information structure, i.e., noisy communication channel ν. An otherwise exceedingly easy problem immediately dismisses any former results in the LQG literature, as vividly pointed out by Witsenhausen [22].

Furthermore, general problems of finding optimal control law or decision rule are still very difficult, theoretically and computationally. These difficulties include: 1) evaluation of expected cost value and gradients, if applicable, by means of Monte Carlo simulations; 2) immense search space caused by, for example, curse of dimensionality; and 3) lack of useful structure which is most exemplified by many combinatorial problems and human made systems (see Ho [12]). In this paper, there is an underlying LQG structure together with the Gaussian channel which facilitates the formulation by step functions, especially in the derivation of $g^*(y|f)$ and $D_f(y)$ [cf. (9) and (11)]. This in turn removes us from the first difficulty because Monte Carlo simulation is replaced by numerical integration of the total cost expression. Nevertheless, we are still faced by the second difficulty—huge search space. This immensity can be seen if one discretizes the domain and range of the function to be optimized, which in our case here is $f(x)$. Suppose the domain and range are discretized to have $|X|$ and $|U|$ points respectively, then the number of possible functions to be searched is $|U|^{|X|}$. For merely $|U| = |X| = 100$, the search space becomes a sizable 100^{100}. Papadimitriou and Tsitsiklis [19] have shown that the discretized version of Witsenhausen problem is NP-complete.[23]

TABLE V
Improvement of Cost Over Time

	Stage 1 Cost	Stage 2 Cost	Total Cost
W [22]	0.404230878394	0.000022320501	0.404253198895
BB [2]	0.363380227653	0.001634601425	0.365014829078
DH [7]	0.139488401953	0.053072896985	0.192561298938
LLH	0.131884081844	0.035429123524	0.167313205368

To deal with such a rapid growth of the search space, use of a hierarchy can be seen very effective. It breaks down a large problem into vertically integrated stages of subproblems. This is synonymous to tactics such as divide-and-conquer or chain of command. In fact, this may be the only possible way to combat combinatorial explosion. Hierarchical search also means coarsening the initial solutions and overviewing the landscape of the solution space. In the context of step functions as in this paper, the coarsening part is the survey of the number of steps. Precisely, let Θ_n be the space of the n-step functions. Then, one can easily see that

$$\Theta_{0.5} \subset \Theta_1 \subset \Theta_{1.5} \subset \Theta_2 \subset \Theta_{2.5} \subset \cdots \subset \Theta_{14} \subset \cdots.$$

By first identifying the subspace in $\Theta_{3.5}$, we have come to understand the trade-offs between the stage one and stage two costs. In this process, we have "learned" more about the problem and its solutions. When further examining the quadratic nature of the cost terms, we are assured that the addition of small segments contributes positively to reducing the total cost.

The application of hierarchical search is very wide and less problem and constraint dependent. It has been demonstrated in other functional optimization and decision problems as well. We have solved a one-dimensional (1-D) problem in this paper. Another example is in stock option pricing (Patsis [20]) which involve finding optimal exercising strategies as a 1-D functional optimization problem. In this piece of work, sampling and search method, and successive narrowing down of search space hierarchically, have been applied very successfully.

More adept readers would by now ponder: How to construct a useful hierarchy for problems in general? Are there any guidelines to help the task?

[21]A performance distribution function tells us what fraction of the space has a cost less than a specified value.

[22]The horizontal axis is scaled as $-\log_{10}(\log_{10}(\text{cost} + 1))$. Whereas the vertical axis is scaled as $-\log_{10}$ (quantile).

[23]Ho and Chu [9] have shown, however, that proper discretization can lead to a convex optimization problem, which, unfortunately, is engulfed by a set of severely imposed constraints.

In light of our experience in this paper, there are three procedural guidelines at work: *goal softening*, *specializing*, and *learning*. These guidelines are useful not only in the difficulty faced in this paper, namely, huge search space, but also when evaluation of cost objective is subject to heavy noise/imprecision, and when the underlying structure of a problem is not apparent—difficulties generally faced by current generation of more complex optimal system problems where "structure" are far less evident than in earlier generation of problems. Typical examples are human made systems such as airport traffic control and internet communication networks.

Goal softening means that instead of asking for the best solution(s), one settles with some "good enough" solutions. The intuition behind this is that by softening the goal one is effectively enlarging the "target" area to a lot other good solutions, which may have much higher chance to be observed. The idea of goal softening belongs to a broader optimization technique called ordinal optimization, first propelled by Ho *et al.* [10]. The importance of goal softening has been explained in Ho [12], Lau and Ho [14], and Lee *et al.* [15]. In Deng and Ho [7], ordinal optimization has been applied very successfully in the Witsenhausen problem to locate the 2-step function in Section I-B. In this paper, despite a very convenient formulation by step functions, we do not insist on searching the 14-step, or higher, step function space in the first place which presumably yields the "best" solution. We have instead coasened the solution space into various subspaces of n-step functions, for $n < 7$, and look for some "good enough" solutions. We have then specialized in the subspace of 3.5-step functions. Specializing is important because it dwells into the more promising regions in the search space and intensifies important attributes of potentially optimal solutions.[24] In our case here, we have improved by the scatter search technique to a local minimum of 3.5-step functions. Lastly, we cannot emphasize any less the learning part about the problem structure with respect to placement of breakpoints and the tradeoffs between the two stages.

VI. Conclusion

In this paper, we have developed an analytical framework using the step function formulation for the Witsenhausen counterexample. This is a two-stage team decision problem that requires minimization of a signaling function $f(x)$. Our formulation has not only speeded up the calculation of the cost objective but has also facilitated learning about the cost terms in the two different stages. We search in a hierarchical fashion on the number of steps first, and narrow down the search in the subspace of 3.5-step functions. Then, by analyzing the respective cost terms in stage one and stage two, and by adding the small segments as discussed in Section III-D, we are able to achieve a 13% improvement from the previously known best (Deng and Ho [7]) and more than 54% better than previous results obtained by other authors for a benchmark instance. The quality of our solution is assessed through order statistics in Section IV-A.

We have also pointed out the importance of hierarchical search in general optimal control and decision problems. In this paper, we only deal with one dimensional functional optimization but higher dimensional problems remain as a challenge. We have outlined three procedural components for general functional optimization and decision problems. These three components, namely, goal softening, specializing and learning are working at best when they are considered mutually, and by no means have to be applied in any particular order. Meanwhile, there are also traits in many computational intelligence techniques such as neural network, fuzzy logic, genetic algorithms and Tabu search, to name a few, to which the three components work in a complementary fashion. The proper combination of these components can help narrowing down the search space and zeroing in the optimum. This is, however, an important research area.

Given our understanding of the Witsenhausen counterexample, and the quality of our solution, this paper offers a challenge to find more new insight for future investigators of the long outstanding problem.

Appendix I
Continuity of J

Here, we are to show the continuity of the cost function J with respect to the parameters of step function, i.e., $b_1, \ldots, b_n$, and $w_1, \ldots, w_n$. Our illustration is based on the following facts: products, sums and differences of continuous functions are continuous, and quotient of continuous functions is also continuous provided that the quotient is defined, i.e., the denominator is nonzero.

Proposition: For given n a finite positive integer and $f(x)$ a n-step function, $J(f)$ is continuous with respect to the parameters $b_1, \ldots, b_n$ and $w_1, \ldots, w_n$.[25]

Proof: Recall from (10) the stage one cost as a function of its parameters

$$k^2E[x - f(x)]^2 = 1 - \frac{2}{\sqrt{2\pi\sigma^2}} \sum_{i=1}^{n} w_i \left[\exp\left(\frac{b_{i+1}^2}{2\sigma^2}\right) - \exp\left(\frac{b_i^2}{2\sigma^2}\right)\right] + 2\sum_{i=1}^{n} p_i k^2 w_i^2.$$

Since

$$p_i = \frac{1}{2}\left(erf\left(\frac{b_{i+1}}{\sqrt{2}\sigma}\right) - erf\left(\frac{b_i}{\sqrt{2}\sigma}\right)\right), \quad i = 1, 2, \ldots, n$$

it is clear then all p_i's are continuous with respect to the parameters b_i and b_{i+1}. It is obvious that the stage one cost is continuous in b_is and w_is. Meanwhile, the stage two cost is $E[(f(x) - g(f(x) + \nu))^2] = 1 - I(D_f)$ where

$$I(D_f) = \int \left(\frac{d}{dy} D_f(y)\right)^2 \frac{dy}{D_f(y)}$$

and $D_f(y)$ is the density function of y. For $f(x)$ a step function, $D_f(y)$ can be rewritten as

$$D_f(y) = 2\sqrt{2\pi}\phi(y) \sum_{i=1}^{n} p_i \phi(w_i) \cosh(w_i y).$$

[24]The No Free Lunch Theorem (Wolpert and MaCready [23]) says that, without structural assumptions learning about these structures and specializing to them, no algorithm can be expected to perform better on the average than the simple blind search.

[25]Note, the "$n.5$-step" functions are in a subclass of $(n + 1)$-step functions with the restriction that the first signaling level being set at $w_1 = 0$.

Let $h(y) = 2\sqrt{2\pi}\sum_{i=1}^{n} p_i\phi(w_i)\cosh(w_i y)$, we can easily see that $h(y)$ is also continuous in b_is and w_is. Moreover, as long as the b_is are not all identical, there exists $p_j > 0$ for some j. As a result, $h(y) > 0$ for all y. Hence, $D_f(y) = \phi(y)h(y) > 0$ for all y. Now

$$\frac{dD_f}{dy} = \phi\frac{dh}{dy} + h\frac{d\phi}{dy}$$

and

$$\frac{dh}{dy} = 2\sqrt{2\pi}\sum_{i=1}^{n} p_i w_i \phi(w_i)\sinh(w_i y).$$

Therefore

$$\begin{aligned}\frac{dD_f}{dy} &= 2\sqrt{2\pi}\phi(y)\sum_{i=1}^{n} p_i w_i \phi(w_i)\sinh(w_i y) \\ &\quad - 2\sqrt{2\pi}y\phi(y)\sum_{i=1}^{n} p_i\phi(w_i)\cosh(w_i y) \\ &= 2\sqrt{2\pi}\phi(y)\left\{\sum_{i=1}^{n} p_i\phi(w_i)[w_i\sinh(w_i y) - y\cosh(w_i y)]\right\}.\end{aligned}$$

We can see that $(dD_f/dy)^2$ is continuous with respect to b_i's and w_i's, and so is $(1/D_f)(dD_f/dy)^2$. Therefore, the integrand of $I(D_f)$ is continuous with respect to b_I's and w_is. Finally, in order that $I(D_f)$ is continuous, we need to see if the integral indeed exists. Clearly, $I(D_f) \geq 0$. By the quadratic nature of the second stage cost, we have $0 \leq E[(f(x) - g(f(x)+\nu))^2] = 1 - I(D_f)$, and, hence, $I(D_f) \leq 1$. Since both stage one and stage two costs are continuous with respect to the parameters, therefore, $J(f)$ is also continuous with respect to the parameters in the step function formulation. □

APPENDIX II
SCATTER SEARCH

Here is a description of the scatter search technique employed in this paper. An n-step function is completely defined by $(n; b_1, \ldots, b_n; w_1, \ldots, w_n)$ as shown in Section II-B-1. By scatter search, we mean to search functions of which one of the parameters (breakpoints and signaling levels) is perturbed by a specific amount from the current iterate. Formally, at iteration t, let $(n; b_1^t, \ldots, b_n^t; w_1^t, \ldots, w_n^t)$ be the current solution and B^t be the set of all neighboring points, i.e.,

$$B^t = \left\{(n; b_1^t, b_2^t \pm r_2\varepsilon, \ldots, b_n^t \pm r_n\varepsilon; w_1^t \pm s_1\varepsilon, \ldots, w_n^t \pm s_n\varepsilon): \sum_{i=2}^{n} r_i + \sum_{j=1}^{n} s_j = 1,\ r_i,\ s_j \in \{0, 1\}\right\}$$

for some appropriate ε which decreases over time.[26] For $t = 1$, the initial solution $(n; b_1^1, \ldots, b_n^1; w_1^1, \ldots, w_n^1)$ is obtained, for example, as in Section III-B. By evaluating the costs for all neighbors in B^t, we move to the one with the lowest cost, i.e., the next iterate $(n; b_1^{t+1}, \ldots, b_n^{t+1}; w_1^{t+1}, \ldots, w_n^{t+1})$ is set to the neighbor in B^t with the lowest cost. The process is done iteratively until all perturbations produce no further improvement.

[26]Note, the parameter b_1 is not perturbed since it is set to 0 due to the symmetry about the origin.

TABLE VI
THE PERFORMANCE VALUE OF THE BEST DESIGN

	f	f^+	f^-
Performance Value	0.238461	0.186649	0.181396

APPENDIX III
ODD FUNCTION ASSUMPTION

The cost term can be broken into stage one and stage two costs. Recall equation (5), stage two cost, $E[(f(x) - g(f(x) + \nu))^2]$, can be re-written as $(1 - I(D_f))$ where $I(D_f)$ is the Fisher information of the random variable y with density $D_f(y)$. The random variable y is the noisy observation made by player two. The Fisher information is defined as

$$I(D_f) = \int \left(\frac{d}{dy} D_f(y)\right)^2 \frac{dy}{D_f(y)}.$$

The Fisher information is a gauge of how easy it is for player two to distinguish between the different signals sent by player one over the noisy channel [16]. Since our goal is to minimize the overall cost term, we need to minimize the stage two cost without compromising the stage one cost. In order to minimize the stage two cost, we have to maximize the Fisher information. Since Fisher information is a gauge of the quality of the signals sent by player one, it does not make a difference if the strategy utilized, $f(x)$ is an odd function or not. Instead, the separation between adjacent signals, for a given noise level, are more important in determining the value of the Fisher information. On the other hand, the stage one cost could only be optimal if $f(x)$ is odd due to the symmetry of the cost term, e.g., $E[k^2(f(x) - x)^2]$. Given that stage one cost requires the optimal strategy to be odd and stage two cost does not care, it is reasonable to assume that the optimal strategy for the total cost would need to be an odd function. Thus, we have restricted our search to odd functions only.

Below are some empirical results which further support our assumption that the optimal $f(x)$ should be an odd function. We randomly generated 50 000 nonodd functions—nonsymmetric functions. We denote them as f's. We construct symmetric odd function from the nonsymmetric f's. In particular, we created symmetric functions by reflecting the portion of $f(x)$ over the positive domain and negative domain about the origin, denote as f^+ and f^-, respectively. In 3363 out of the 50 000 instances, we observe the nonsymmetric functions outperforming their symmetric counterparts. This translates to less than 7% of the time that the nonsymmetric functions are better. Moreover, we have found that the best among the nonsymmetric one is outperformed by the best among the symmetric one's by a large margin. In other words, if we are seeking the functions with the best performance values, out of the 150 000 functions (fs, f^+s and f^-s) that we have evaluated, we should look at the one's that are odd. The values are listed below in Table VI.

This supports the idea that the very best designs would be an odd function.

ACKNOWLEDGMENT

The first author would like to thank H. McLaughlin for providing the proof that a step function can approximate any given function under certain regularity conditions.

REFERENCES

[1] M. Baglietto, T. Parisini, and R. Zoppoli, "Nonlinear approximations for the solution of team optimal control problems," in *Proc. CDC*, vol. 5, San Diego, 1997, pp. 4592–4594.
[2] R. Banal and T. Basar, "Stochastic teams with nonclassical information revisited: When is an affine law optimal," *IEEE Trans. Automat. Contr.*, vol. 32, pp. 554–559, June 1987.
[3] A. E. Bryson and Y.-C. Ho, *Applied Optimal Control*: Hemisphere, 1975.
[4] M. L. Cohen, "The fisher information and convexity," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 591–592, 1968.
[5] L.-Y. Dai and C.-H. Chen, "Rate of convergence for ordinal comparison of dependent simulations in discrete event dynamic systems," *J. Optim. Theory Appl.*, vol. 94, no. 1, pp. 29–54, July 1997.
[6] H. A. David, *Order Statistics*. New York: Wiley, 1970.
[7] M. Deng and Y.-C. Ho, "Sampling-selection method for stochastic optimization problems," *Automatica*, vol. 35, no. 2, pp. 331–338, 1999.
[8] Y.-C. Ho and T. S. Chang, "Another look at the nonclassical information structure problem," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 537–540, June 1980.
[9] Y.-C. Ho and K.-C. Chu, "Team decision theory and information structures in optimal control problems–Part I," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 15–22, Feb. 1972.
[10] Y.-C. Ho, R. S. Sreenivas, and P. Vakili, "Ordinal optimization in DEDS," *J. Discrete Event Dyna. Syst.*, vol. 2, no. 1, pp. 61–88, 1992.
[11] Y.-C. Ho and M. E. Larson, "Ordinal optimization approach to rare event probability problems," *J. Discrete Event Dyna. Syst.*, vol. 5, no. 2/3, pp. 281–301, 1995.
[12] Y.-C. Ho, "Heuristics, rules of thumb, and the 80/20 proposition," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 1025–1027, May 1994.
[13] ——, "A future for control systems research: A vision of what, how and why," unpublished, 1998.
[14] T. W. E. Lau and Y.-C. Ho, "Alignment probabilities and subset selection in ordinal optimization," *J. Optim. Appl.*, vol. 93, no. 3, pp. 455–489, 1997.
[15] L. H. Lee, T. W. E. Lau, and Y.-C. Ho, "Explanation of goal softening in ordinal optimization," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 94–99, Jan. 1999.
[16] L. H. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.
[17] W.-G. Li, "Vector and constraint ordinal optimization—Theory and practice," Ph.D. dissertation, Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA, Jan. 1998.
[18] S. Mitter and A. Sahai, "Information and control: Witsenhausen revisited," M.I.T., LIDS Laboratory, report, 1998.
[19] C. H. Papadimitriou and J. N. Tsitsiklis, "Intractable problems in control theory," *SIAM J. Control Optim.*, vol. 24, no. 4, pp. 639–654, 1986.
[20] N. T. Patsis, "Pricing American-style exotic options using ordinal optimization," Ph.D. dissertation, Harvard University, Cambridge, MA, 1998.
[21] P. Vakili, L. Mollamustaflaglu, and Y.-C. Ho, "Massively parallel simulation of a class of discrete event systems," in *Proc. IEEE Frontiers MPC Symp.*, Washington, D.C., 1992.
[22] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM J. Control*, vol. 6, no. 1, pp. 131–147, 1968.
[23] D. G. Woplert and W. G. Macready, "No free lunch theorems for search," *IEEE Trans. Evol. Comput.*, vol. 1, pp. 67–82, Apr. 1997.
[24] M. S. Yang, L. H. Lee, and Y.-C. Ho, "On stochastic optimization and its applications to manufacturing," in *Proc. AMS-SIAM Summer Seminar The Mathematics of Stochastic Manufacturing Systems*, Williamsburg, VA, 1996.
[25] M. S. Yang, "Ordinal optimization and its application to complex deterministic problems," Ph.D. dissertation, Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA, Nov. 1997.
[26] M. Baglietto, T. Parisini, and R. Zoppoli, "Numerical solutions to the Wittenhausen counterexample by approximating networks," IEEE Trans. Automat. Contr., to be published.

Jonathan T. Lee is currently a Ph.D. student in the Division of Engineering and Applied Sciences at Harvard University, Cambridge, MA.

His research interests include simulation, analysis and optimization of complex systems with applications to manufacturing systems, inventory systems, and communication networks.

Edward Lau received the B.S. degree in manufacturing engineering from Boston University, Boston, MA, where he was Chu Scholar, and the Ph.D. degree in applied mathematics from Harvard University, Cambridge, MA.

His research interests include systems simulation, optimization, and decision sciences. He has also performed high level tool design and applications for industrial and commercial enterprises.

Yu-Chi (Larry) Ho (S'54– M'55– SM'62– F'73–LF '97) received the B.S. and M.S. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, and the Ph.D. degree in applied mathematics from Harvard University, Cambridge, MA.

Except for three years of full-time industrial work, he has been on the Harvard faculty where he is the T. Jefferson Coolidge Chair in Applied Mathematics and Gordon McKay Professor of Systems Engineering. He was also the Visiting Professor to the Cockrell Family Regents Chair in Engineering at the University of Texas, Austin, in 1989. He has published over 140 articles and three books, one of which (co-authored with A. E. Bryson, Jr.) has been translated into both Russian and Chinese and made the list of Citation Classics as one of the most referenced works on the subject of optimal control. He is on the editorial boards of several international journals, and is the Editor-in-Chief of the *International Journal on Discrete Event Dynamic Systems*. His research interests lie at the intersection of control system theory, operations research, and computational intelligence. He has contributed to topics ranging from optimal control, differential games, information structure, multiperson decision analysis, to incentive control, and, since 1983, exclusively to discrete-event dynamic systems, perturbation analysis, ordinal optimization, and computational intelligence.

Dr. Ho has been the recipient of various fellowships and awards, including the Gugganheim (1970) and the IEEE Field Award for Control Engineering and Science (1989), the Chiang Technology Achievement Prize (1993), the Bellam Control Heritage Award (1999) of the Automatic Control Council, and the AMSE Rufus Oldenberg Award (1999). He is a Distinguished Member of the Control Systems Society, and was elected a member of the U.S. National Academy of Engineering in 1987, a foreign member of the Chinese Academy of Sciences and the Chinese Academy of Engineering in 2000. In addition to service on various governmental and industrial panels and professional society administrative bodies, he was President of the IEEE Robotics and Automation Society in 1998 and cofounder of Network Dynamics, Inc., a software firm specializing in industrial automation.