Optimal Designs for Gamut Models

William D Heavlin, Google, Inc.

September 2018

Abstract

Experiments exploring directions of steepest ascent (DSAs) are isomorphic to one-factor-at-a-time (1AAT) experiments — they try runs along the DSA at varying distances, typically with intermediate factor levels (IFLs). A natural extension of the 1AAT DSA approach likewise encourages IFLs, incorporates the DSA, and includes also the factor × DSA interactions. When combined with locally weighted scatterplot smoothing ("lowess"), this defines a class of varying-coefficient models called gamut models. Our main optimal design criterion is average ("integrated") mean square prediction error (I-optimality), which is relatively friendly to IFLs. We adapt this criterion to the lowess-based varying-coefficient gamut models and assess the resulting designs.

keywords: gamuts, I-optimality, lowess, optimal design, varying-coefficient models.

1 Approaches to Experimental Design

At least since Fisher (1935), factorial designs have formed a cornerstone of experimental design theory. The key properties of factorial designs are (1) their high efficiency in terms of sample size, (2) their enhanced sensitivity to detecting interactions, and (3) their capacity to enclose rather high-dimensional factor spaces.

Box and Hunter (1961a, 1961b) brought two-level experimental designs to substantial maturity, culminating in the classic texts Box, Hunter, and Hunter (2005) and Montgomery (2013). This line of practice favors symmetry properties and highlights orthogonal arrays. This paradigm is also tolerant of model ambiguity, e.g. resolution IV designs, which confound multiple second-order interactions with one another.

A complementary approach formulates experimental designs as solutions to optimization problems. This algorithmic paradigm poses an objective function, scores tentative solutions, and makes iterative refinements. Unlike the symmetry approaches just noted, this so-called *optimal design* paradigm presumes a uniquely specified model, and some of the tension between these two paradigms revolves around the relevance of resolution IV designs.

For a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the least squares estimate of $\boldsymbol{\beta}$ is $\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top y$ and the variance-covariance matrix of $\mathbf{b}$ is proportional to $(\mathbf{X}^\top\mathbf{X})^{-1}$. Wald (1943) introduced D-optimality, which minimizes the volume of the least squares coefficients ellipsoid, proportional to a *determinant*: $\texttt{det}((\mathbf{X}^\top\mathbf{X})^{-1}) = 1/\texttt{det}(\mathbf{X}^\top\mathbf{X})$. An alternative criterion, G-optimality, minimizes the largest (*greatest*) prediction error in the design space: $\texttt{max}(\mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x})$. Kiefer and Wolfowitz (1960) demonstrate the asymptotic equivalence of D- and G-optimality.

Like G-optimality, Studden (1977) focused on prediction error. However, rather than minimizing its worst or largest value, Studden proposes averaging it — *integrating* it — over the factor space: $\int \mathbf{x}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}\,d\mathbf{x} = \texttt{trace}((\mathbf{X}^\top\mathbf{X})^{-1}\int \mathbf{x}\mathbf{x}^\top d\mathbf{x})$; this criterion he terms I-optimality. Note how the I-optimal criterion consists of two factors, the design, represented by the term $(\mathbf{X}^\top\mathbf{X})^{-1}$ and the underlying factor space, represented by the covariance-like second moment matrix $\int \mathbf{x}\mathbf{x}^\top d\mathbf{x}$.

From examples Hardin and Sloane (1993) conclude that I-optimal designs are distinct from D- and G-optimal designs: The I-optimal criterion distinguishes between designs with the same D-optimal score; further, the I-optimal criterion gives rather more preference to points near the center of the design region.

Didactically, the teaching of experimental design usually requires both displacing a practice and replacing it with a new concept. The practice to be displaced consists of an approach often assimilated in middle and high school science classes: (a) run a control group first, (b) hold everything constant but one variable, which is systematically varied, then (c) repeat (b) with a different variable.

Such a process consists of several distinct ideas: (0) the role of control runs, (1) varying only one variable at a time ("1AAT"), and (2) setting that selected variable to multiple different values — intermediate factor levels or "IFLs." In contrast, the experimental design theory that replaces this high-school style assesses designs numerically, and then demonstrates the statistical inefficiency and insensitivity to factor interactions of 1AAT and IFLs. (We put aside the issue of control runs, which is rather more nuanced.) These results are then used to build out a statistically based experimental design paradigm.

2 Steepest Ascent

The didactically standard template for statistical experimental design proposes this series of steps: (0) a first phase of establishing stability, often by repeated control runs, (1) screening designs, typically two-level and low-resolution, sufficient for a linear model, (2) follow-up designs that assess and perhaps resolve two-factor interactions, and (3) supplemental runs that estimate quadratic effects.

In the context of optimization, Box, Hunter, and Hunter (2005, chapter 12) and Montgomery (2013, chapter 11) suggest gradient ascent. The prescription essentially explores along a single direction: $\mathbf{x}_0 + \lambda_i \mathbf{g}$, for $0 < \lambda_1 < \lambda_2 < \cdots$, where $\mathbf{x}_0$ denotes the current design center, $\mathbf{g}$ the estimated gradient at $\mathbf{x}_0$, and $\{\lambda_1, \lambda_2, ...\}$ denote various scalars guiding the magnitude of $\mathbf{g}$. The statistical literature calls $\mathbf{g}$ the *direction of steepest ascent* (DSA). In the standard template described above, the DSA can be computed from data arising in either step (1) or step (2) and consists of the derivative of the response surface.

In the numerical analysis literature, gradient ascent (or descent) is central. Levenberg (1944) and Marquardt (1963) are jointly credited for the algorithm defining a damped hessian-based update. Press et al (1989) names the hessian-free quasi-Newton version BFGS, for the near-concurrent contributions of Broyden (1970), Fletcher (1970), Goldfarb (1970), and Shanno (1970). Dekker (1969) and Brent (1973) are known for the iterative exploring for and bounding of a solution along a defined direction.

For optimizing response surfaces, Box, Hunter, and Hunter (2005, chapter 12) and Montgomery (2013, chapter 11) present the gradient ascent methods from the deterministic numerical analysis literature without modification. However, in the context of experimental design, both the DSA is estimated with uncertainty and any attempted new setting is measured with experimental error. At any rate, their proposal for a Brent-method style of one-dimensional exploration resembles strongly a 1AAT experiment, somewhat inconsistent with the standard statistical pedagogy that favors factorial experiments.

3 Interactions and Steepest Ascent

A natural expansion of the 1AAT form of DSA would incorporate interactions; this corresponds to the response surface's gradient or DSA $\mathbf{g}(\mathbf{x})$ changing as a function of the coordinates $\mathbf{x}$.

Consider the n-vector $\mathbf{g}$ and the $n \times J$ matrix $\mathbf{X}$. Define the direct product $\mathbf{g} \otimes \mathbf{X}$ as giving an $n \times J$ matrix whose i-th row is $(g_i X_{i1}, g_i X_{i2}, g_i X_{i3}, \ldots, g_i X_{iJ})$. A DSA model with interactions has this expectation:

$$\mathbb{E}\{\mathbf{y}|\mathbf{g}, \mathbf{X}\} = \beta_0 + \mathbf{g}\beta_1 + \mathbf{g} \otimes \mathbf{X}\boldsymbol{\beta}. \tag{1}$$

$\boldsymbol{\beta} = 0$ corresponds to no interactions, and in this case the DSA direction $\mathbf{g}$ is sufficient.

Given a DSA estimate $\mathbf{g}$, model (1) enables an optimal design approach for choosing additional coordinates. In form, (1) resembles one of the models associated with resolution IV designs, that which estimates all interactions with one particular factor. In this case, (1) poses all interactions with the feature (and gamut) $\mathbf{g}$.

4 Gamut Models

Writing of deterministic models, Constantine et al. (2016) assert that many physical systems have a ridge function structure like (1). In the context of large scale data analysis, Heavlin (2016) postulates something similar, "gamut models," about which we now elaborate.

A gamut model is a form of varying-coefficient model, in the sense of Hastie and Tibshirani (1993), where the coefficients vary as a function of a particular variable, the gamut. Thus, its expectation is

$$\mathbb{E}\{y_i|g_i,\mathbf{x}_i\} = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}(g_i), \text{where } g_i = g(\mathbf{x}_i) \tag{2}$$

The gradient at coordinate $\mathbf{x}$ is $\boldsymbol{\beta}(g(\mathbf{x}))$, varying as a function of the scalar-valued gamut function $g(\cdot)$. Model (2) is not identifiable, but Heavlin (2016) overcomes this with three ideas: (a) Find a gamut of convenience that can be computed on all n observations. (b) Define a vector of n weights $\mathbf{w}(g)$ that give more weight to observations with gamut values near g and less on those farther away. (c) Estimate the coefficients $\boldsymbol{\beta}(g)$ by weighted least squares with weights $\mathbf{w}(g)$.

The simpler implementations of gamut models make pragmatic choices: Gamut (a) can be taken to be the predicted values of an unweighted least squares model. Weights (b) are derived from the tricubic lowess (locally weighted scatterplot smoother) weights, as originally proposed by Cleveland (1979). To be invariant to monotone transformations, gamut values are sometimes replaced by their ranks.

Recall the key relations for weighted least squares: Near a particular scalar gamut value g, define n weights $\mathbf{w}(g)$ and their corresponding diagonal matrix $\mathbf{D}_g = \texttt{diag}(\mathbf{w}(g))$. Then, the coefficients $\mathbf{b}(g)$ can be calculated as

$$\mathbf{b}(g) = (\mathbf{X}^\top \mathbf{D}_g \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}_g \mathbf{y} \tag{3}$$

with this associated variance-covariance matrix:

$$\mathbb{V}_g \equiv \mathbb{COV}\{\mathbf{b}(g)\} = \sigma^2 (\mathbf{X}^\top \mathbf{D}_g \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}_g^2 \mathbf{X} (\mathbf{X}^\top \mathbf{D}_g \mathbf{X})^{-1}, \tag{4}$$

where σ^2 is the usual variance of the underlying experimental error, and typically estimated by the residual mean square error. Note that (4) is scale-invariant: $\mathbb{V}_g$ remains unchanged if $\mathbf{D}_g$ is replaced by $a\mathbf{D}_g$ for any positive scalar a.

5 Gamuts and Optimal Design

We emphasize that the coefficients estimated by (3) need only be useful in a certain part of the factor space, in the neighborhood where in some sense the gamut function $g(\mathbf{x}) \approx g_0$, for some scalar g_0. One can reasonably represent this neighborhood, which we have not been defined precisely, by postulating a probability density function (pdf) $p_g(\mathbf{x})$ on the $\mathbf{x}$-domain and factor space.

Let us define $\boldsymbol{\Omega}(g) \equiv \int \mathbf{x}\mathbf{x}^\top \, p_g(\mathbf{x}) d\mathbf{x}$. For a series of scalar gamut values $g_1, g_2, \ldots$ we likewise have a series of $\mathbf{x}$-domain second-moment matrices $\boldsymbol{\Omega}(g_1)$, $\boldsymbol{\Omega}(g_2)$, $\ldots$.

A natural extension of I-optimality in the case of gamut models is this lower-is-better criterion:

$$\sum_k \texttt{trace}(\mathbb{V}_{g_k} \boldsymbol{\Omega}(g_k)) \tag{5}$$

(5) seeks to minimize the sum of weighted average prediction error over regions, mediated by pdfs $p_g(\mathbf{x})$, where their predictions are likely to be relevant. Note that corresponding extensions to D- or G-optimality are less straightforward: D-optimality focuses on how well the coefficients are estimated, but without the context of the factor space. G-optimality, in contrast, focuses on the largest prediction error over the factor space; it is not obvious how to restrict G's worst-case construct to a subspace of interest. In this sense, I-optimality adapts well to gamut models while D- and G-optimality would seem to adapt awkwardly at best.

As presented, criterion (5) minimizes prediction error, so promotes exploration. Note that the original motivation for calculating DSAs is to identify better operating points, exploitation — in other words, a DSA points toward higher utility, and gamut functions measure this utility. A natural modification therefore would

$$\texttt{minimize} \sum_k \texttt{trace}(\mathbb{V}_{g_k} \boldsymbol{\Omega}(g_k)) - \alpha \sum_i g(\mathbf{x_i}) \tag{6}$$

for some scalar $\alpha \geq 0$. When α is zero, (5) is equivalent to (6); when $\alpha > 0$, some consideration is given to higher utility gamut values.

6 An Example

We illustrate these ideas by a toy example. Suppose we have four factors and our gamut consists of their sum: $g(\mathbf{x}) = \mathbf{x}^\top \mathbf{1}$. We seek a 5-run design sensitive to the interactions with this direction. If we put no utility on the gamut itself, we can use criterion (5) or equivalently criterion (6) with $\alpha = 0$.

The resulting design is presented in Table 2. This design has two low-gamut design points and three high-gamut points. Observe how the three high-gamut points represent 1AAT variation from the $\mathbf{x} = (1, 1, 1, 1)$ maximum gamut point — which, we emphasize, is not run. In this regard, it is reminiscent of the resolution III design 2^{3-1}_{III} — sometimes called L4 or OA(4) — presented in Table 1. In the case of Table 2, of course, the 1AAT portion moves to middling values like -0.15.

Consider now the solution when we use criterion (6) with $\alpha = 0.5$. Because of the extra utility associated with the gamut values, the resulting design, presented in Table 3, now has one low-gamut and four high-gamut values, the low-gamut point is no longer so low, and the remaining four runs all represent 1AAT trials from the maximum gamut point $(1, 1, 1, 1)$ — which we emphasize again is not run. If anything, the resemblance to the resolution III in Table 1 has grown stronger. Now, the 1AAT portion has slightly more de-tuned values like -0.23. The average gamut value has increased, while the four high-gamut values recoup some of the information needed to estimate interactions with the gamut.

X01	X02	X03
-1	-1	-1
-1	1	1
1	-1	1
1	1	-1

Table 1: The design variously called 2^{3-1}_{III}, L4, and OA(4)

X01	X02	X03	X04	gamut
-1.00	-1.00	-0.92	-1.00	-3.92
0.87	-1.00	-1.00	-1.00	-2.13
1.00	1.00	1.00	-0.15	2.85
1.00	-0.15	1.00	1.00	2.85
1.00	1.00	-0.04	1.00	2.96

Table 2: 5-run explore design in four factors, $\alpha = 0$

X01	X02	X03	X04	gamut
-0.54	-0.81	-0.76	-0.66	-2.77
1.00	1.00	-0.26	1.00	2.74
-0.23	1.00	1.00	1.00	2.77
1.00	1.00	1.00	-0.23	2.77
1.00	-0.21	1.00	1.00	2.79

Table 3: 5-run explore/exploit design in four factors, $\alpha = 0.5$

7 Conclusion

The I-optimal criterion consists of two factors, one holding the information of the experimental design and another describing the second moments of the factor space. The partitioning of these two factors facilitates building designs for the class of varying-coefficient models we call here gamut models.

The resulting designs have a coarse structure and a fine-grained structure. The coarse structure shows some intermediate factor levels (IFLs), and both low-gamut and high-gamut values. The latter point helps estimate the coefficient associated with the gamut direction itself, β_1 in (1).

As to the fine-grained structure, one end of the gamut range are nearly equal gamut values achieved by distinct factor settings. These same-gamut, different-settings runs share a common pattern: Relative to an (absent) highest-gamut run, each factor is de-tuned one at a time from its optimum. The average of these help estimate β_1 in (1); the variation among them helps estimate the interactions of each factor with the gamut direction, $\boldsymbol{\beta}$ in (1).

8 References

Box, G.E.P. and Hunter, J.S. (1961a) The 2^{k-p} fractional factorial designs, Pt. I. *Technometrics*, 3(3): 311-351. (1961b) Pt. II. 3(4): 449-458, 1961.

Box, G.E.P., Hunter, J.S., and Hunter, W.G. (2005), *Statistics for Experimenters: design, discovery, and innovation.* New York: Wiley.

Brent, RP (1973), *Algorithms for Minimization Without Derivatives.* Prentice-Hall.

Broyden, C. G. (1970), The convergence of a class of double-rank minimization algorithms, *Journal of the Institute of Mathematics and Its Applications*, 6: 76–90.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, 74: 829-836.

Constantine, P.G., del Rosario, Z., and Iaccarino, G. (2106). Many physical laws are ridge functions, arxiv:1605.07974.

Dekker, T.J. (1969), Finding a zero by means of successive linear interpolation, in *Constructive Aspects of the Fundamental Theorem of Algebra*, Dejon, B. and Henrici, P., eds. Wiley.

Fisher, R.A. (1971) [1935]. *The Design of Experiments*, 9th edition, Macmillan.

Fletcher, R. (1970), A new approach to variable metric algorithms, *Computer Journal*, 13(3): 317–322.

Goldfarb, D. (1970), A family of variable metric updates derived by variational means, *Mathematics of Computation*, 24(109): 23–26

Hardin, R.H. and Sloane, N.J.A. (1993). A new approach to the construction of optimal designs, *Journal of Statistical Planning and Inference*, 37(3), 339-369,

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models (with discussion), *Journal of the Royal Statistical Society*, Series B, 55: 757-796.

Heavlin, W.D. (2016). Modeling with gamuts, *ASA Proceedings.*

Kiefer, J.C. and Wolfowitz, J. (1960). The equivalence of two extremum problems, *Canadian Journal of Mathematics*, 12, 363-366.

Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least squares. *Quarterly of Applied Mathematics.* 2: 164–168

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics.* 11(2): 431–441.

Montgomery, D.C. (2013). *Design and Analysis of Experiments*, New York: Wiley.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1989). *Numerical Recipes*, New York: Cambridge University Press.

Shanno, D.F. (1970), Conditioning of quasi-Newton methods for function minimization, *Mathematics of Computation*, 24 (111): 647–656

Studden, W. J. (1977) Optimal designs for integrated variance in polynomial regression, in *Statistical Decision Theory and Related Topics* II, eds. S. S. Gupta and D. S. Moore, New York: Academic Press.

Wald, A. (1943) On the efficient design of statistical investigations. *Annals of Mathematical Statistics* 14(2), 134–140.