

# Optimization Problems

The goal in an optimization problem is to find the point at which the minimum (or maximum) of a real, scalar function  $f$  occurs and, usually, to find the value of the function at that point.

We use the term “optimum” or “extremum” to refer to a minimum or maximum. We commonly consider the minimization problem only.

Maximizing  $f(x)$  is equivalent to minimizing its negative,  $-f(x)$ .

The general unconstrained optimization problem can be stated as the problem of finding the vector  $x_*$  where the minimum occurs, or of finding the minimum value of the function

$$\min_x f(x) = f(x_*).$$

The function  $f$  is called the *objective function*. The elements of  $x$  are often called *decision variables*.

# Statistical Methods as Optimization Problems

Many common statistical methods are developed as solutions to optimization problems.

In a common method of statistical estimation, we maximize a likelihood, which is a function proportional to a probability density at the point of the observed data.

In another method of estimation and in standard modeling techniques, we minimize a norm of the residuals. The best fit of a model is often defined in terms of a minimum of a norm, such as least squares.

Other uses of optimization in statistical applications occur prior to collection of data, for example, when we design an experiment or a survey so as to minimize experimental or sampling errors.

When a statistical method is based on the solution of an optimization problem, to formulate that problem unambiguously helps us both to understand the method and to decide whether the method is appropriate to the purposes for which it is applied.

# Fitting Statistical Models

A common type of model relates one variable to others in a form such as

$$y = f(x; \theta) + \epsilon,$$

in which  $\theta$  is a vector of *fixed parameters* with unknown and unobservable values.

*Fitting* the model is mechanically equivalent to *estimating*  $\theta$ .

The most familiar form of this model is the linear model

$$y = x^T \beta + \epsilon,$$

where  $x$  and  $\beta$  are vectors. We also often assume that  $\epsilon$  is a random variable with a normal distribution.

There are two approaches, both optimization problems.

In both, the first step is to replace the fixed unknown parameter with a variable.

## Fitted Residuals

For a given value of the variable used in place of the parameter and for each pair of observed values  $(y_i, x_i)$ , we form a *fitted residual*. In the case of the linear model, for example, we form the fitted residual

$$r_i(b) = y_i - x_i^T b,$$

in terms of a variable  $b$  in place of the unknown estimand  $\beta$ .

We should note the distinction between the fitted residuals  $r_i(b)$  and the unobservable residuals or errors,  $\epsilon_i = y_i - x_i^T \beta$ . When  $\epsilon$  is a random variable,  $\epsilon_i$  is a realization of the random variable, but the fitted residual  $r_i(b)$  *is not* a realization of  $\epsilon$  (as beginning statistics students sometimes think of it).

Our objective in fitting the model is to choose a value of the variable, which when used in place of the parameter, minimizes the fitted residuals.

We note that the residuals, either fitted or those from the “true” model, are vertical distances.

# Minimizing Residuals

The idea of minimizing the residuals from the observed data in the model is intuitively appealing.

Because there is a residual at each observation, however, “minimizing the residuals” is not well-defined without additional statements. When there are several things to be minimized, we must decide on some way of combining them into a single measure. It is then the single measure that we seek to minimize.

A useful type of overall measure is a *norm* of the vector of residuals. The measure, of course, is a function of  $b$ . The most obvious measure, perhaps, may just be the sum of the absolute values:

$$R_1(b) = \sum_{i=1}^n |y_i - x_i^T b|.$$

This quantity is called the  $L_1$  *norm* of the vector of residuals  $r(b)$ , and is denoted as  $\|r(b)\|_1$ . Another possible measure is the sum of the squares:

$$R_2(b) = \sum_{i=1}^n |y_i - x_i^T b|^2.$$

This quantity is called the  $L_2$  *norm* of the vector of residuals  $r(b)$ , and is denoted as  $\|r(b)\|_2$ .

# Least Squares of Residuals

For various reasons, the most common approach to fit the model with the given data is *least squares*; that is, to use  $R_2(\mathbf{b})$  as the overall measure of the residuals to minimize. With  $n$  observations, the ordinary least squares estimator of  $\beta$  in the linear model is the solution to the *optimization problem*

$$\min_{\mathbf{b}} \sum_{i=1}^n (y_i - x_i^T \mathbf{b})^2.$$

This optimization problem is relatively simple, and its solution can be expressed in a closed form as a system of linear equations.

Sometimes we may know that  $\beta$  must satisfy certain restrictions, and so we modify the optimization problem to impose constraints. For the case of  $\beta \geq 0$ , for example, we formulate the nonnegative least squares *optimization problem*

$$\min_{\mathbf{b} \geq 0} \sum_{i=1}^n (y_i - x_i^T \mathbf{b})^2.$$

This optimization problem is considerably more complicated than the unconstrained problem. Its solution cannot be expressed in a closed form.

# Nonlinear Least Squares

When the original model is nonlinear, again, we form a norm of the residuals, but the least squares problem is much more difficult both computationally and conceptually than the linear least squares problem.

$$\min_t \sum_{i=1}^n (y_i - f(x_i; t))^2$$

In general, there is no closed-form solution to this optimization problem.

# Minimizing Other Functions of the Residuals

For the general objective of minimizing the residuals we have alternatives. We may measure the overall size of the residuals by

$$R_\rho(t) = \sum_{i=1}^n \rho(y_i - f(x_i; t)),$$

where  $\rho(\cdot)$  is some function of  $r_i = y_i - f(x_i; t)$ .

Instead of minimizing the sum of the squares of the residuals, we fit the model by minimizing this measure; that is, by solving an *optimization problem* such as

$$\min_t \sum_{i=1}^n \rho(y_i - f(x_i; t)). \quad (1)$$

Depending on  $\rho(\cdot)$ , this problem is much more difficult both computationally and conceptually than the least squares problem, in which  $\rho(r) = r^2$ . One common choice of  $\rho$  is just the absolute value itself, and the problem of fitting the model is the *optimization problem*

$$\min_t \sum_{i=1}^n |y_i - f(x_i; t)|. \quad (2)$$

There is no closed-form solution to this simple least-absolute-values problem, even in the linear case.

## Treating Residuals Differently

In addition to choosing a function of the individual  $r_i$ , we might also reconsider how we choose to combine several individual residual values into a single measure. We may want to treat some residuals differently from others, resulting in the *optimization problem*

$$\min_t \sum_{i=1}^n w(y_i, x_i, t) \rho(y_i - f(x_i; t)),$$

where  $w(y_i, x_i, t)$  is a nonnegative function. Because in practice, for this minimization problem, it is usually not explicitly a function of  $y_i$ ,  $x_i$ , and  $t$ , we often write  $w(y_i, x_i, t)$  as a simple fixed weight,  $w_i$ .

A common instance is the weighted linear least squares problem with fixed weights, in which the function to be minimized is

$$\sum_{i=1}^n w_i (y_i - x_i^T b)^2.$$

The weights do not materially change the complexity of this problem. It has a closed-form solution, just as the unweighted (or equally-weighted) problem.

## Regularization of the Solution

We may also regularize the minimum residuals problem with additional criteria. We form a weighted linear combination of two functions of  $t$ ,

$$\sum_{i=1}^n w(y_i, x_i, t) \rho(y_i - f(x_i; t)) + \lambda g(t),$$

where  $g$  is some nonnegative function and  $\lambda$  is some nonnegative scalar used to tune the optimization problem. The simplest instance of this kind of regularization is in ridge regression with a linear model, in which  $w$  is constant,  $\rho(z) = z^2$ ,  $f(x_i; b) = x_i^T b$ , and  $g(b) = b^T b$ . The *optimization problem* is

$$\min_b \left( \sum_{i=1}^n (y_i - x_i^T b)^2 + \lambda b^T b \right).$$

In ridge regression, we minimize a weighted combination of  $L_2$  norms of the vector of residuals,  $r(b)$ , and of the coefficients,  $b$ . In lasso regression with a linear model, an  $L_2$  norm is applied to the residuals and an  $L_1$  norm is applied to the coefficients, and the *optimization problem* is

$$\min_b (\|r(b)\|_2 + \lambda \|b\|_1).$$

# Minimizing Residuals Nonparametrically

There are other ways of approaching the problem of fitting the model.

Instead of fixing the form of the function  $f$  and determining a suitable value of  $\theta$ , we may assume the form of  $f$  is unknown (or uninteresting) and approximate it in a way that the approximation  $\tilde{f}(x)$  fits the data closely. This kind of approach is *nonparametric*.

The optimization problem that involves determination of  $\tilde{f}(x)$  is a quite different problem from our previous examples. In any case, however, the first step is to be clear about our objective.

Just to minimize some function of the residuals is not sufficient. Unless we add some conditions on  $\tilde{f}(x)$ , there are infinitely many solutions that yield 0 residuals (assuming no two observations on  $x$  have the same value).

# Minimizing a Regularized Function of the Residuals

In a nonparametric approach, in addition to a requirement that the residuals be small, we may regularize the problem with other criteria for fitting the model.

For example, we may require that our approximation  $\tilde{f}(x)$  be twice-differentiable and be “smooth”. If we measure the roughness or non-smoothness of a twice-differentiable function  $f$  over a domain  $D$  by the integral of the square of the second derivative,

$$\mathcal{R}_{22}(f) = \int_D (f''(x))^2 dx,$$

we can include this expression in our optimization problem.

Our overall optimization would be a weighted combination of this expression and some measure of the residuals. In regularized least squares, the *optimization problem* is

$$\min_{\tilde{f}} \sum_{i=1}^n (y_i - \tilde{f}(x_i))^2 + \lambda \mathcal{R}_{22}(\tilde{f}),$$

with the restriction that  $\tilde{f}$  be twice-differentiable. In this formulation,  $\lambda$  is a nonnegative smoothing parameter.

# Maximum Likelihood Estimation

Another way of fitting the model  $y = f(x; \theta) + \epsilon$  is by *maximizing the likelihood function* that arises from the probability distribution of  $\epsilon$ . Given the data, this is the *optimization problem*

$$\max_t \prod_{i=1}^n p(y_i - f(x_i; t)),$$

where  $p(\cdot)$  is the probability function or the probability density function of the random error.

This yields a maximum likelihood estimator (MLE).

Optimization problems of this type can be quite formidable computationally.

For a given probability density  $p(\cdot)$ , the maximization problem for MLE may be equivalent to a minimization problem of fitting residuals.

# Regularization in Maximum Likelihood Estimation

In a nonparametric formulation of the maximum likelihood approach, we are faced with the same kind of indeterminate problem as in a nonparametric approach to minimizing residuals.

Unless we regularize the problem with additional criteria, the problem is ill-posed. We can “penalize” the likelihood with a regularization measure that decreases (remember we are maximizing) as we move away from the desirable solution.

For example, if we require that the function be smooth (and, hence, twice-differentiable), we may form the *optimization problem*

$$\max_{\tilde{f}} \prod_{i=1}^n p(y_i - \tilde{f}(x_i)) e^{-\lambda \mathcal{R}_{22}(\tilde{f})},$$

where  $\mathcal{R}_{22}(\cdot)$  is the same functional as before.

# Experimental Design

Other optimization problems in statistics arise in optimal design of experiments and in the construction of optimal sampling plans. In design of experiments, we often assume a linear relationship between  $y$  and an  $m$ -vector  $x$ , and we anticipate collecting  $n$  observations,  $(y_i, x_i)$ , into an  $n$ -vector  $y$  and an  $n \times m$  matrix  $X$ . We may express the relationship as

$$y = \beta_0 1 + X\beta + \epsilon.$$

Under the assumption that the residuals are independently distributed with a constant variance,  $\sigma^2$ , the variance-covariance matrix of estimable linear functions of the least squares solution are formed from

$$(X^T X)^{-1} \sigma^2.$$

Because we may be able to choose the settings at which we make observations, we may attempt to choose them in such a way as to minimize variances of certain estimators. The variance of a particular estimator is minimized by maximizing some function of  $X^T X$ . A common method in experimental design results in the *optimization problem*

$$\max_{\text{all factor settings}} \det(X^T X).$$

# Optimal Sampling Design

In designing a sampling plan, we have the problem of allocating the sample sizes  $n_h$  within strata. For given population strata sizes  $N_h$  and known (or assumed) within-strata variances  $v_h$ , the *optimization problem* has the form

$$\min_{1 \leq n_h \leq N_h} \sum_h N_h \left( \frac{N_h}{n_h} - 1 \right) v_h.$$

(Often the lower bound on  $n_h$  is taken to be 2.)

Determination of optimal sampling plans can be a very complicated if more complicated designs are used or if other considerations, such as cost of conducting the sample, are brought to bear. In a two-stage design, for example, we have the problem of allocating the sample sizes  $n_h$  and  $m_h$  within various strata and across multiple stages.

For given population sizes  $N_h$  and known within-strata variances for the first and second stages  $v_h$  and  $v_{2h}$ , we have the *optimization problem*

$$\min_{n_h, m_h} \left( \sum_h N_h \left( \frac{N_h}{n_h} - 1 \right) v_h + \sum_h \frac{N_h^2}{n_h m_h} v_{2h} \right).$$

# Imputation, Editing, and Calibration

Often data collected as responses to questionnaires contain obvious inaccuracies, and the sampling agency wants to correct those inaccuracies while changing the data as little as possible.

Often a list of edits is available that contains rules that data items must obey.

The techniques of data editing are also employed in record matching. In this application, the amount of change required to make an identifying set of fields in one record to correspond to those fields in another record is assessed. If no change or only a small change is required, the two records are deemed to match.

A common instance of data adjustment, usually called “calibration” rather than “editing”, is the adjustment of tabular data to fixed marginal totals.

# Imputation, Editing, and Calibration

In sampling applications, an observational record may consist of many individual responses, and certain information may already be available for some items. For a given item,  $X$ , a total,  $\tau_{X_D}$ , over a domain,  $D$ , may be known.

If the observations on  $X$  in the domain are  $x_1, \dots, x_n$ , and the corresponding sampling weights are  $d_1, \dots, d_n$ , the estimated domain total is  $\hat{\tau}_{X_D} = \sum d_k x_k$ . The calibration problem is to choose new weights  $w_1, \dots, w_n$ , “close to” the original weight, but such that  $\sum w_k x_k = \tau_{X_D}$ . The differences in  $w_k$  and  $d_k$  can be measured in terms of  $|w_k - d_k|$  or  $w_k/d_k$ . (All weights are nonzero.) If  $g(\cdot)$  is a measure of the distance from  $w_k$  to  $d_k$  measured in terms of  $w_k/d_k$  we form the calibration *optimization problem* as

$$\min_{w_k} \sum_{i=1}^n d_k g(w_k/d_k),$$

subject to the requirement  $\sum w_k x_k = \tau_{X_D}$ . The terms in the function to be minimized are weighted by the original weights, which generally correspond to the importance attached to the individual observations.

# Orthogonal Residuals

An interesting problem arises in regression when the model accommodates observational or sampling errors in  $x$ . The errors-in-variables problem may lead to an orthogonal regression approach in which the residuals are not measured in a vertical direction, as in the usual formulation, or in a horizontal direction, as in a calibration problem, but rather in an orthogonal (or normal) direction from the model surface.

For orthogonal residuals, because the direction of the residuals depends on the fitted model, we cannot express the optimization problem in terms of a vector norm.

# Orthogonal Residuals

For a linear model of the form

$$y \approx X\beta,$$

the problem of finding  $b$  so as to minimize the  $L_2$  norm of the orthogonal residuals is the problem of finding  $b$  that satisfies

$$\tilde{y} = \tilde{X}b$$

where  $\tilde{y}$  and  $\tilde{X}$  are solutions to the *optimization problem*

$$\min_{\tilde{y} \in \text{span}(\tilde{X})} \left\| \begin{bmatrix} X & y \end{bmatrix} - \begin{bmatrix} \tilde{X} & \tilde{y} \end{bmatrix} \right\|_{\text{F}},$$

where  $\text{span}(\tilde{X})$  is the column space of  $\tilde{X}$ ,  $\begin{bmatrix} X & y \end{bmatrix}$  and  $\begin{bmatrix} \tilde{X} & \tilde{y} \end{bmatrix}$  are the matrices formed by adjoining as the last column the vector in the second position to the matrix in the first position, and  $\|\cdot\|_{\text{F}}$  is the Frobenius matrix norm.

To minimize other norms of the orthogonal residuals requires solution of even more complicated optimization problems. Even minimizing the  $L_2$  norm of the orthogonal residuals, as above, cannot be performed by evaluating a closed-form expression.

# Response Surface Methodology

In an important class of applications, a model is used to represent the effects of some controllable factors measured by  $x$  on the response of some quantity of interest measured by  $y$ .

The objective in *response surface methods* is to determine the optimal combination of settings of  $x$  for the response  $y$ ; that is, to solve the *optimization problem*

$$\max_x f(x; \theta).$$

This optimization problem cannot be solved because  $\theta$  is unknown; hence, response surface methodology actually involves two optimization problems: fitting of the model (which is usually just a trivial application of least squares), that is, determination of the “best” value of  $t$  in place of  $\theta$ , and then finding the optimum of the function  $f(x; t)$  over  $x$ .

# Estimation of Functions

An interesting statistical problem is the estimation of a function that relates one variable to the expected value of another variable,  $E(Y) = f(x)$ .

In some cases we can write this model as:

$$Y = f(x) + \epsilon.$$

In this version, the difference  $Y - f(x)$  is modeled as the random variable  $\epsilon$ . There are other possible formulations, but we will consider only this additive model. The problem is to estimate  $f$  given a set of observations  $y_i$  and  $x_i$ .

We have a general nonparametric approach to this problem with a regularized optimization problem for a solution.

# ML Estimation of Functions

An alternative is maximum likelihood estimation, which of course requires a probability distribution.

One common method in function estimation is to define a set of *basis functions* that span some class of functions that either contains  $f$ .

After a set of basis functions,  $\{q_k\}$ , is chosen, our estimating function  $h$  is chosen as

$$h(x) = \sum_k c_k q_k(x),$$

where the  $c_k$  are taken as a solution to the *optimization problem*

$$\min_{c_k} \|f - c_k q_k\|.$$

The basis functions are often chosen as orthogonal polynomials.

One of the most interesting problems of function estimation in statistics is nonparametric probability density estimation.

# Clustering and Classification

Less formal statistical methods also use optimization. In k-means clustering, for example, we seek a partition of a dataset into a preset number of groups  $k$  that minimizes the variation within each group. Each variable may have a different variation, of course. The variation of the  $j^{\text{th}}$  variable in the  $g^{\text{th}}$  group is measured by the within sum-of-squares:

$$s_{j(g)}^2 = \frac{\sum_{i=1}^{n_g} (x_{ij(g)} - \bar{x}_{j(g)})^2}{n_g - 1},$$

where  $n_g$  is the number of observations in the  $g^{\text{th}}$  group, and  $\bar{x}_{j(g)}$  is the mean of the  $j^{\text{th}}$  variable in the  $g^{\text{th}}$  group. For data with  $m$  variables there are  $m$  such quantities. In k-means clustering the *optimization problem* is

$$\min_{\text{all partitions}} \sum_{g=1}^k \sum_{j=1}^m s_{j(g)}^2.$$

When groups or classes are known, the problem of determining to which group a given observation belongs is called “classification”. In classification, we seek to determine optimal discriminators to define class membership.

# Formulation of an Optimization Problem

The formulation of a statistical problem or any problem in data analysis as an optimization often helps us to understand the problem and to focus our efforts on the relevant aspects of the problem.

In calibration of tables or in data editing, for example, we seek adjustments that represent minimal changes from the original data.

If we do not think clearly about the problem, the resulting optimization problem may not be *well-posed*; that is, it may not have an unambiguous solution. Functional optimization problems often are not well-posed without the regularization component.

One of the most worrisome problems arises when the optimization problem has multiple points of optimality. The presence of multiple local minima should cause us to think about the problem more deeply.

The objective function should correspond to the objectives of the analysis.

# Formulation of an Optimization Problem

In formulating a statistical problem as an optimization problem, we must be careful not to change the statistical objectives. The objective function and the constraints should reflect the desired statistical methodology.

An example in the literature of how available software can cause the analyst to reformulate the objective function began with the problem of fitting a linear regression model with linear constraints; that is, a problem in which the constraints on  $b$  were of the form  $Ab \leq c$ .

It turns out that an optimization problem of *least absolute values* regression, with  $f(x_i; b) = x_i^T b$ , that is, *linear* regression, and with constraints of the form  $Ab \leq c$ , can be formulated as a linear programming problem with some additional constraints (see Charnes, Cooper, and Ferguson, 1955), and solved easily using available software. At the time, there was no readily available software for constrained least squares regression, so the reformulated problem was solved.

# Formulation of an Optimization Problem

The solution to an optimization problem is in some sense “best” for that problem and its objective function. This fact may mean that the solution is considerably less good for some other optimization problem. It is often the case, therefore, that an optimal solution is not robust to assumptions about the phenomenon being studied.

Use of optimization methods is likely to magnify the effects of any assumptions.

# Optimization of Multiple Objectives

In most practical applications with optimization, there are more than one objective.

A simple example of this is the general optimization problem of minimizing the residuals in order to fit a model. Because there are many residuals, we must decide how we want to minimize them all simultaneously. Of course, the obvious solution to this quandary is just to minimize the sum (of some function) of the residuals.

Even in this simple example, however, there may be reasons to combine the residuals differentially, as in weighted regression.

# Optimization of Multiple Objectives

In a problem of optimal sampling design the general objective is to minimize the variance of an estimator. Realistically, there are many estimates that result from a single survey; there are several attributes (estimands) crossed with several strata.

The problem of how to address the problem of minimizing the variances of all within strata estimators in a single sampling design requires consideration of the relative importance of the estimates, any constraints on variances and/or coefficients of variation, and constraints on the cost of the survey. It may also be possible to relax hard constraints on variances or costs and to include them as additional objectives.

There are various ways of accommodating multiple objectives and constraints. The simplest, of course, is to form a weighted sum. Constraints can be incorporated as a weighted component of the objective function. The weight controls the extent to which the constraint is met.

In most cases, a certain amount of interaction between the decision maker and the optimization procedure is required.

# Solution of an Optimization Problem and Solution of the Problem of Interest

We must be careful to formulate an optimization problem that captures the objective of the real problem.

By the nature of optimization, the effect of any assumptions about the real problem may be magnified. Therefore, in addition to being concerned about the agreement of the formulation of the optimization problem with the problem of interest, we must also be careful about the analysis of that underlying problem.

Even when the problem is formulated correctly, there may be difficulties in using the optimization problem in a statistical method. A simple example of this occurs in using elementary calculus in maximum likelihood estimation. We may have local maxima or other stationary points, or the solution may be on the boundary.

There are many other problems that can arise in using optimization methods. For example, there may be more than one point of optimality. This kind of problem brings into question the formulation of the statistical method as an optimization problem.

# No Free Lunches

Given a general problem such as to solve the linear system,

$$Ax = b,$$

we would like to have a general-purpose program that we can call on to find the solution.

In the case of linear systems, we can almost do this. The only really hard thing may be whether the problem is a system of Diophantine equations (and how often have you heard of *that?*). Assuming on restriction on the solution space, the only relevant questions may be whether or not the system is extremely large, and if it is, whether or not it is sparse, whether or not we want extremely high accuracy, and what to do in case there is either no solution or more than one solution.

The situation is very different for a general problem such the optimization problem,

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & x \in S. \end{aligned}$$

The Diophantine-type situation occurs very often, so that's the first big choice.

There are other problems, however.

# No Free Lunches

In the optimization problem,

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & x \in S, \end{aligned}$$

all of the issues of the linear system arise, and they're generally harder to address.

The other type of problem arises from the range of possibilities for the function  $f(x)$ .

The classic example is where  $f(x) = g(x) + a\delta(x - x_0)$ , where  $g(\cdot)$  is relatively nice, and  $\delta(\cdot)$  is the Dirac delta. This problem cannot be solved by any general optimization method unless  $a$  and  $x_0$  are nice.

Wolpert and Macready (1997) state the problem formally and in a more general way. (“No free lunch theorems for the maximum of an unknown function”, *IEEE Transactions on Evolutionary Computation* **1**, 67–82.)

The bottom line is that we should not seek a general-purpose optimization program.