4 Quadratics

The objective functions used in linear least-squares and regularized least-squares are multi-dimensional quadratics. We now analyze multi-dimensional quadratics further. We will see many more uses of quadratics later in the course, particularly when dealing with Gaussian distributions.

The general form of a quadratic of one variable is given by:

$$f(x) \;=\; w_2 x^2 + w_1 x + w_0 \;. \tag{1}$$

This can also be written in a slightly different way (called standard form):

$$f(x) \;=\; a(x-b)^2 + c \;, \tag{2}$$

where $a = w_2, b = -w_1/(2w_2)$, and $c = w_0 - w_1^2/4w_2$. These two forms are equivalent, and it is easy to go back and forth between them (e.g., given a, b, c, what are w_0, w_1, w_2?). In the latter form, it is easy to visualize the shape of the curve. It is a bowl with a minimum (or maximum) at b, and the "width" of the bowl is determined by the magnitude of a. The sign of a tells us which direction the bowl points (a positive means a convex bowl, a negative means a concave bowl), and c tells us how high or low the bowl goes (at $x = b$). We will now generalize these intuitions for higher-dimensional quadratics.

The general form for a 2D quadratic function is

$$f(x_1, x_2) \;=\; w_{1,1} x_1^2 + w_{1,2} x_1 x_2 + w_{2,2} x_2^2 + w_1 x_1 + w_2 x_2 + w_0 \;, \tag{3}$$

and, for an N-D quadratic, it is

$$f(x_1, ...x_N) \;=\; \sum_{1 \le i \le N, 1 \le j \le N} w_{i,j} x_i x_j + \sum_{1 \le i \le N} w_i x_i + w_0 \;. \tag{4}$$

Note that there are three sets of terms here, namely, the quadratic terms ($\sum w_{i,j} x_i x_j$), the linear terms ($\sum w_i x_i$) and the constant term (w_0).

Dealing with these summations is rather cumbersome. We can simplify things by using matrix-vector notation. Let $\mathbf{x}$ be an N-dimensional column vector, written $\mathbf{x} = [x_1, ...x_N]^T$. Then we can write a quadratic as

$$f(\mathbf{x}) \;=\; \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c \;, \tag{5}$$

where

$$\mathbf{A} \;=\; \begin{bmatrix} w_{1,1} & \dots & w_{1,N} \\ \vdots & w_{i,j} & \vdots \\ w_{N,1} & \dots & w_{N,N} \end{bmatrix} \tag{6}$$

$$\mathbf{b} \;=\; [w_1, ..., w_N]^T \tag{7}$$

$$c \;=\; w_0 \tag{8}$$

You should verify for yourself that these different forms are equivalent, by multiplying out all the elements of $f(\mathbf{x})$, either in the 2D case or, using summations, the general $N - D$ case.

For many of the manipulations we'll need later, it is helpful for $\mathbf{A}$ to be symmetric, i.e., to have $w_{i,j} = w_{j,i}$. In fact, it should be clear that these off-diagonal entries are redundant. So, if we are a given a quadratic for which $\mathbf{A}$ is not symmetric, we can symmetrize it in the following way;

$$f(\mathbf{x}) \;=\; \mathbf{x}^T(\frac{1}{2}(\mathbf{A}+\mathbf{A}^T))\mathbf{x} + \mathbf{b}^T\mathbf{x} + c \;=\; \mathbf{x}^T\tilde{\mathbf{A}}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c\,, \tag{9}$$

and use $\tilde{\mathbf{A}} = \frac{1}{2}(\mathbf{A}+\mathbf{A}^T)$ instead. You should confirm for yourself that this is equivalent to the original quadratic.

As before, we can convert the quadratic to a form that provides an intuitive interpretation:

$$f(\mathbf{x}) \;=\; (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) + d\,, \tag{10}$$

where $\boldsymbol{\mu} = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}, d = c - \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$, assuming that $\mathbf{A}^{-1}$ exists. Note the similarity here to the 1-D case. As before, this function is a bowl-shape in N dimensions, with curvature specified by the matrix $\mathbf{A}$, and with a single stationary point $\boldsymbol{\mu}$.[1] However, fully understanding the shape of $f(\mathbf{x})$ is a bit more subtle and interesting.

4.1 Optimizing a Quadratic

Suppose we wish to find the stationary points (minimum or maximum) of a quadratic

$$f(\mathbf{x}) \;=\; \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c\,. \tag{11}$$

The stationary points occur where all partial derivatives are zero, i.e., $\partial f/\partial x_i = 0$ for all i. The gradient of a function is the vector comprising the partial derivatives of the function, i.e.,

$$\nabla f \;\equiv\; \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial N}\right]^T\,. \tag{12}$$

At stationary points it must therefore be true that $\nabla f = [0,\ldots,0]^T$. Let us assume that $\mathbf{A}$ is symmetric (if it's not, then we can symmetrize it as above). Equation (11) is a very common form of cost function (e.g., the log probability of a Gaussian distribution as we will later see), and so the form of its gradient is important to examine.

Due to the linearity of the differentiation operator, we can look at each of the three terms of Eq. (11) separately. The last (constant) term does not depend on $\mathbf{x}$ and so we can ignore it (because its derivative is zero). Let us examine the first term. If we write out the individual terms within the

[1] A stationary (or critical) point means a setting of $\mathbf{x}$ where the gradient is zero.

vectors/matrices, we get:

$$(x_1 \dots x_N) \begin{pmatrix} a_{11} & \dots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \dots & a_{NN} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \tag{13}$$

$$= (x_1 a_{11} + x_2 a_{21} + \dots + x_N a_{N1} x_1 a_{12} + x_2 a_{22} + \dots \tag{14}$$

$$\dots + x_1 a_{1N} + x_2 a_{2N} + \dots + x_N a_{NN}) \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \tag{15}$$

$$= x_1^2 a_{11} + x_1 x_2 a_{21} + \dots + x_1 x_N a_{N1} + x_1 x_2 a_{12} + x_2^2 a_{22} + \dots + x_N x_2 a_{N2} + \dots \tag{16}$$

$$\dots x_1 x_N a_{1N} + x_2 x_N a_{2N} + \dots + x_N^2 a_{NN} \tag{17}$$

$$= \sum_{ij} a_{ij} x_i x_j \tag{18}$$

The i^{th} element of the gradient corresponds to $\partial f / \partial x_i$. So in the expression above, for the terms in the gradient corresponding to each x_i, we only need to consider the terms involving x_i (others will have derivative zero), namely

$$x_i^2 a_{ii} + \sum_{j \neq i} x_i x_j (a_{ij} + a_{ji}) \tag{19}$$

The gradient then has a very simple form:

$$\frac{\partial \left(\mathbf{x}^T \mathbf{A} \mathbf{x} \right)}{\partial x_i} \;=\; 2 x_i a_{ii} + \sum_{j \neq i} x_j (a_{ij} + a_{ji}). \tag{20}$$

We can write a single expression for all of the x_i using matrix/vector form:

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} \;=\; (\mathbf{A} + \mathbf{A}^T) \mathbf{x}. \tag{21}$$

You should multiply this out for yourself to see that this corresponds to the individual terms above. If $\mathbf{A}$ is symmetric, then we have

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} \;=\; 2 \mathbf{A} \mathbf{x}. \tag{22}$$

This is also a very helpful rule that you should remember. The next term in the cost function, $\mathbf{b}^T \mathbf{x}$, has an even simpler gradient. Note that this is simply a dot product, and the result is a scalar:

$$\mathbf{b}^T \mathbf{x} \;=\; b_1 x_1 + b_2 x_2 + \dots + b_N x_N \; . \tag{23}$$

Only one term corresponds to each x_i, so $\partial f / \partial x_i = b_i$. We can again express this in matrix/vector form:

$$\frac{\partial \left(\mathbf{b}^T \mathbf{x} \right)}{\partial \mathbf{x}} \;=\; \mathbf{b} \; . \tag{24}$$

This is another helpful rule that you will encounter again. If we use both of the expressions we have just derived, and set the gradient of the cost function to zero, we get

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \;=\; 2\mathbf{A}\mathbf{x} + \mathbf{b} \;=\; [0, \ldots, 0]^T \;. \tag{25}$$

The optimum is given by the solution to this system of equations (called *normal equations*):

$$\mathbf{x} \;=\; -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b} \tag{26}$$

In the case of scalar x, this reduces to $x = -b/2a$. For linear regression with multi-dimensional inputs above (see Eqn. (**??**)): $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ and $\mathbf{b} = -2\mathbf{X}^T\mathbf{y}$. As an exercise, convince yourself that this is true.

