Optimization from Calculus

Chris Cornwell

September 11 and 16, 2025

Outline

Outline

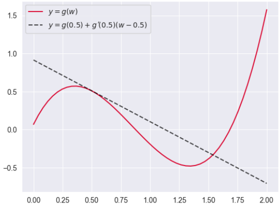
Linear approximations

$g : \mathbb{R} \to \mathbb{R}$, a twice-differentiable function (at least). Write w for input to g, so $g(w)$.

Linear approximation to $g(w)$. At a point $(v, g(v))$ on its graph, function whose graph is the tangent line:

$$h(w) = g(v) + g'(v)(w - v).$$

- Also called first order Taylor series approximation (or Taylor polynomial) of g.
- Approximates values of g, for inputs near v.



Second order (quadratic) approximations

To approximate g better (and in a larger interval around v), can use the second order Taylor polynomial. This is the function:

$$h(w) = g(v) + g'(v)(w - v) + \frac{1}{2}g''(v)(w - v)^2.$$

Second order (quadratic) approximations

To approximate g better (and in a larger interval around v), can use the second order Taylor polynomial. This is the function:

$$h(w) = g(v) + g'(v)(w - v) + \frac{1}{2}g''(v)(w - v)^2.$$

- Incorporates both first and second derivative.

Second order (quadratic) approximations

To approximate g better (and in a larger interval around v), can use the second order Taylor polynomial. This is the function:

$$h(w) = g(v) + g'(v)(w - v) + \frac{1}{2} g''(v)(w - v)^2.$$

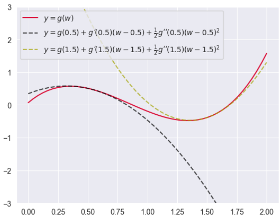
- Incorporates both first and second derivative.
- **Local minimum.** If $g'(v) = 0$ and $g''(v) > 0$, then $h(w) \geq g(v)$. (near v, approximation is good, values of $g(w) \geq g(v)$).

Second order (quadratic) approximations

To approximate g better (and in a larger interval around v), can use the second order Taylor polynomial. This is the function:

$$h(w) = g(v) + g'(v)(w - v) + \frac{1}{2}g''(v)(w - v)^2.$$

- Incorporates both first and second derivative.
- **Local minimum.** If $g'(v) = 0$ and $g''(v) > 0$, then $h(w) \geq g(v)$. (near v, approximation is good, values of $g(w) \geq g(v)$).
- If $g'(v) = 0$ and $g''(v) < 0$, then opposite of the last item is true.



First order approximation in multiple variables

Extending to the setting of multiple variables.

- Use the gradient where, if $\mathbf{w} = [w_1 \, w_2 \; \ldots \; w_N]^T$, then

$$\nabla g = [\frac{\partial g}{\partial w_1} \, \frac{\partial g}{\partial w_2} \; \cdots \; \frac{\partial g}{\partial w_N}]^T.$$

First order approximation in multiple variables

Extending to the setting of multiple variables.

- Use the gradient where, if $\mathbf{w} = [w_1 \, w_2 \; \dots \; w_N]^T$, then

$$\nabla g = [\frac{\partial g}{\partial w_1} \; \frac{\partial g}{\partial w_2} \; \cdots \; \frac{\partial g}{\partial w_N}]^T.$$

Linear approximation: is $h(\mathbf{w}) = g(\mathbf{v}) + \nabla g(\mathbf{v})^T(\mathbf{w} - \mathbf{v})$.

First order approximation in multiple variables

Extending to the setting of multiple variables.

- Use the gradient where, if $\mathbf{w} = \left[w_1\ w_2\ \ldots\ w_N\right]^T$, then

$$\nabla g = [\frac{\partial g}{\partial w_1}\ \frac{\partial g}{\partial w_2}\ \cdots\ \frac{\partial g}{\partial w_N}]^T.$$

Linear approximation: is $h(\mathbf{w}) = g(\mathbf{v}) + \nabla g(\mathbf{v})^T(\mathbf{w} - \mathbf{v})$.

- *Is* a (affine) linear function; graph is translation of subspace of $\mathbb{R}^{N+1}$ that is normal to $\left[-\nabla g(\mathbf{v})^T,\ 1\right]^T$

First order approximation in multiple variables

Extending to the setting of multiple variables.

- Use the gradient where, if $\mathbf{w} = [w_1\ w_2\ \ldots\ w_N]^T$, then

$$\nabla g = [\frac{\partial g}{\partial w_1}\ \frac{\partial g}{\partial w_2}\ \cdots\ \frac{\partial g}{\partial w_N}]^T.$$

Linear approximation: is $h(\mathbf{w}) = g(\mathbf{v}) + \nabla g(\mathbf{v})^T(\mathbf{w} - \mathbf{v})$.

- *Is* a (affine) linear function; graph is translation of subspace of $\mathbb{R}^{N+1}$ that is normal to $[-\nabla g(\mathbf{v})^T,\ 1]^T$
- The graph of $h(\mathbf{w})$ is the tangent (hyper)plane to graph of $g(\mathbf{w})$ at the point $(\mathbf{v}, g(\mathbf{v}))$.

Second order approximation in multiple variables

For second order approximations, use a matrix called the **Hessian**, $\nabla^2 g$.

Second order approximation in multiple variables

For second order approximations, use a matrix called the **Hessian,** $\nabla^2 g$.

- The matrix $\nabla^2 g$ (evaluated at $\mathbf{v}$) is the matrix of second order partial derivatives of g:

$$\nabla^2 g = \begin{bmatrix} \frac{\partial^2 g}{\partial w_1 \partial w_1} & \frac{\partial^2 g}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 g}{\partial w_1 \partial w_N} \\ \frac{\partial^2 g}{\partial w_2 \partial w_1} & \frac{\partial^2 g}{\partial w_2 \partial w_2} & \cdots & \frac{\partial^2 g}{\partial w_2 \partial w_N} \\ & & \ddots & \\ \frac{\partial^2 g}{\partial w_N \partial w_1} & \frac{\partial^2 g}{\partial w_N \partial w_2} & \cdots & \frac{\partial^2 g}{\partial w_N \partial w_N} \end{bmatrix}.$$

Second order approximation in multiple variables

For second order approximations, use a matrix called the **Hessian,** $\nabla^2 g$.

- The matrix $\nabla^2 g$ (evaluated at $\mathbf{v}$) is the matrix of second order partial derivatives of g:

$$\nabla^2 g = \begin{bmatrix} \frac{\partial^2 g}{\partial w_1 \partial w_1} & \frac{\partial^2 g}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 g}{\partial w_1 \partial w_N} \\ \frac{\partial^2 g}{\partial w_2 \partial w_1} & \frac{\partial^2 g}{\partial w_2 \partial w_2} & \cdots & \frac{\partial^2 g}{\partial w_2 \partial w_N} \\ & & \ddots & \\ \frac{\partial^2 g}{\partial w_N \partial w_1} & \frac{\partial^2 g}{\partial w_N \partial w_2} & \cdots & \frac{\partial^2 g}{\partial w_N \partial w_N} \end{bmatrix}.$$

Second order approximation: is

$$h(\mathbf{w}) = g(\mathbf{v}) + \nabla g(\mathbf{v})^T(\mathbf{w} - \mathbf{v}) + \frac{1}{2}(\mathbf{w} - \mathbf{v})^T \nabla^2 g(\mathbf{v})(\mathbf{w} - \mathbf{v}).$$

Second order approximation in multiple variables

For second order approximations, use a matrix called the **Hessian,** $\nabla^2 g$.

- The matrix $\nabla^2 g$ (evaluated at $\mathbf{v}$) is the matrix of second order partial derivatives of g:

$$\nabla^2 g = \begin{bmatrix} \frac{\partial^2 g}{\partial w_1 \partial w_1} & \frac{\partial^2 g}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 g}{\partial w_1 \partial w_N} \\ \frac{\partial^2 g}{\partial w_2 \partial w_1} & \frac{\partial^2 g}{\partial w_2 \partial w_2} & \cdots & \frac{\partial^2 g}{\partial w_2 \partial w_N} \\ & & \ddots & \\ \frac{\partial^2 g}{\partial w_N \partial w_1} & \frac{\partial^2 g}{\partial w_N \partial w_2} & \cdots & \frac{\partial^2 g}{\partial w_N \partial w_N} \end{bmatrix}.$$

Second order approximation: is

$$h(\mathbf{w}) = g(\mathbf{v}) + \nabla g(\mathbf{v})^T(\mathbf{w} - \mathbf{v}) + \frac{1}{2}(\mathbf{w} - \mathbf{v})^T \nabla^2 g(\mathbf{v})(\mathbf{w} - \mathbf{v}).$$

- There is something like the second derivative test in this general case; is more involved to describe. In practice, techniques using first order approximations are the most commonly used.

Outline

Stationary points versus (local) minimal

A **stationary point** (or, **critical point**) of g is a point $\mathbf{v}$ where $\nabla g(\mathbf{v})$ is the zero vector. (Called the "first order condition for optimality.")

Stationary points versus (local) minimal

A **stationary point** (or, **critical point**) of g is a point $\mathbf{v}$ where $\nabla g(\mathbf{v})$ is the zero vector. (Called the “first order condition for optimality.”)

- Means that $\mathbf{v}$ is point where all partial derivatives of g are zero.

Stationary points versus (local) minimal

A **stationary point** (or, **critical point**) of g is a point $\mathbf{v}$ where $\nabla g(\mathbf{v})$ is the zero vector. (Called the "first order condition for optimality.")

- Means that $\mathbf{v}$ is point where all partial derivatives of g are zero.
- A minimum of g can only occur at a stationary point. However, other things can happen at a stationary point too – maximum of g or a "saddle" point.

The Chain rule

The **chain rule** is very useful for understanding derivatives and gradients. The more general version of it ("suped up" from Calculus I):

[1]"Df" and "Dg" mean, take the appropriate version of the derivative for the function. If the function is from $\mathbb{R}$ to $\mathbb{R}$, this is the f' from Calc I; if it is multi-variable (from $\mathbb{R}^N$ to $\mathbb{R}$, for some $N > 1$), then Df means ∇f. If f has vector output, there is a matrix of partial derivatives for Df.

The Chain rule

The **chain rule** is very useful for understanding derivatives and gradients. The more general version of it ("suped up" from Calculus I):

Chain rule.
If the composition $f(g(\mathbf{w}))$ is defined and each of f and g are differentiable, then the derivative of the composition is $Df(g(\mathbf{w})) \cdot Dg(\mathbf{w})$.[1]

[1] "Df" and "Dg" mean, take the appropriate version of the derivative for the function. If the function is from $\mathbb{R}$ to $\mathbb{R}$, this is the f' from Calc I; if it is multi-variable (from $\mathbb{R}^N$ to $\mathbb{R}$, for some $N > 1$), then Df means ∇f. If f has vector output, there is a matrix of partial derivatives for Df.

The Chain rule

The **chain rule** is very useful for understanding derivatives and gradients. The more general version of it ("suped up" from Calculus I):

Chain rule.
If the composition $f(g(\mathbf{w}))$ is defined and each of f and g are differentiable, then the derivative of the composition is $Df(g(\mathbf{w})) \cdot Dg(\mathbf{w})$.[1]

Example: given a fixed vector $\mathbf{a}$, set $h(\mathbf{w}) = \frac{1}{1+e^{\mathbf{a}^T\mathbf{w}}}$. This is a composition of $f(x) = \frac{1}{1+e^x}$ and $g(\mathbf{w}) = \mathbf{a}^T\mathbf{w}$. Thus,

$$\nabla h(\mathbf{w}) = f'(g(\mathbf{w})) \cdot \nabla g(\mathbf{w}) = \left(\frac{-e^{\mathbf{a}^T\mathbf{w}}}{1+e^{\mathbf{a}^T\mathbf{w}}}\right)\mathbf{a}.$$

[1] "Df" and "Dg" mean, take the appropriate version of the derivative for the function. If the function is from $\mathbb{R}$ to $\mathbb{R}$, this is the f' from Calc I; if it is multi-variable (from $\mathbb{R}^N$ to $\mathbb{R}$, for some $N > 1$), then Df means ∇f. If f has vector output, there is a matrix of partial derivatives for Df.

Example of stationary points

Here we look at the stationary points of the function $f(w_1, w_2) = w_1^3 + 3w_1w_2 + w_2^3$.

- $\nabla f(w_1, w_2) = \left[3w_1^2 + 3w_2, \quad 3w_1 + 3w_2^2\right]^T$

Example of stationary points

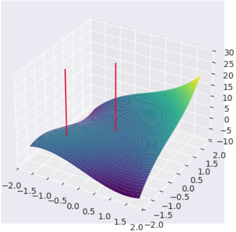
Here we look at the stationary points of the function
$f(w_1, w_2) = w_1^3 + 3w_1 w_2 + w_2^3$.

- $\nabla f(w_1, w_2) = \left[3w_1^2 + 3w_2, \quad 3w_1 + 3w_2^2\right]^T$
- If a stationary pt.,
  $3w_1^2 + 3w_2 = 0$ and
  $3w_1 + 3w_2^2 = 0$; so, $w_2 = -w_1^2$;
  using this and simplifying,
  $w_1 + w_1^4 = 0$.

Example of stationary points

Here we look at the stationary points of the function $f(w_1, w_2) = w_1^3 + 3w_1w_2 + w_2^3$.

- $\nabla f(w_1, w_2) = \left[3w_1^2 + 3w_2, \quad 3w_1 + 3w_2^2\right]^T$
- If a stationary pt., $3w_1^2 + 3w_2 = 0$ and $3w_1 + 3w_2^2 = 0$; so, $w_2 = -w_1^2$; using this and simplifying, $w_1 + w_1^4 = 0$.
- We get: either $w_1 = 0$ (with $w_2 = 0$ also), or $w_1 = -1$ (with $w_2 = -1$ also). See Figure on right.



Vertical lines placed over stationary points $(0, 0)$ and $(-1, -1)$

Convex functions

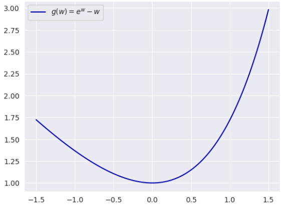
Function $g : \mathbb{R} \to \mathbb{R}$ is called **convex** if $g''(v) \geq 0$ for all v.[2]

[2]A different definition (which is equivalent): g is **convex at** v if the linear approximation at v has a graph that is below the graph of g, near v; it is then **convex** if it is convex at v for all v in $\mathbb{R}$. This extends to multiple variables.

Convex functions

Function $g : \mathbb{R} \to \mathbb{R}$ is called **convex** if $g''(v) \geq 0$ for all v.[2]

Example.
The function $g(w) = e^w - w$ is a convex function, since $g''(w) = e^w$, and e^v is bigger than zero for all v.



[2]A different definition (which is equivalent): g is **convex at** v if the linear approximation at v has a graph that is below the graph of g, near v; it is then **convex** if it is convex at v for all v in $\mathbb{R}$. This extends to multiple variables.