Convex Optimization

DS-GA 1013 / MATH-GA 2824 Optimization-based Data Analysis

http://www.cims.nyu.edu/~cfgranda/pages/OBDA_fall17/index.html

Carlos Fernandez-Granda

Convexity

Differentiable convex functions

Minimizing differentiable convex functions

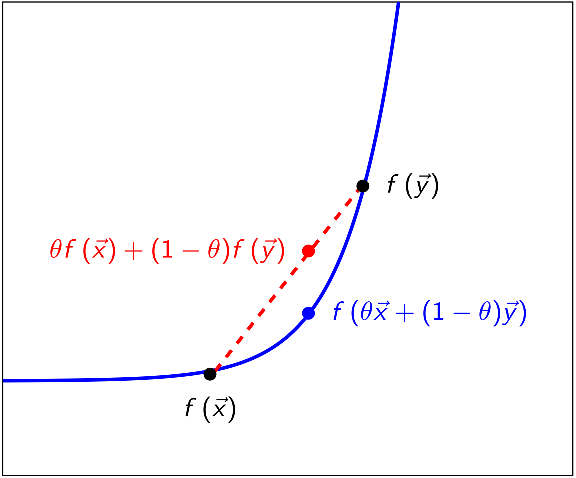
Convex functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$\theta f(\vec{x}) + (1-\theta) f(\vec{y}) \geq f(\theta \vec{x} + (1-\theta)\vec{y})$$

A function f if concave is $-f$ is convex

Convex functions



Linear functions are convex

If f is linear

$$f\left(\theta \vec{x} + (1-\theta)\,\vec{y}\right)$$

Linear functions are convex

If f is linear

$$f\left(\theta\vec{x}+(1-\theta)\,\vec{y}\right)=\theta f\left(\vec{x}\right)+(1-\theta)\,f\left(\vec{y}\right)$$

Strictly convex functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if for any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$\theta f(\vec{x}) + (1-\theta) f(\vec{y}) > f(\theta \vec{x} + (1-\theta)\vec{y})$$

Local minima are global

Any local minimum of a convex function is also a global minimum

Proof

Let $\vec{x}_{\text{loc}}$ be a local minimum: for all $\vec{x} \in \mathbb{R}^n$ such that $||\vec{x} - \vec{x}_{\text{loc}}||_2 \leq \gamma$

$$f\left(\vec{x}_{\text{loc}}\right) \leq f\left(\vec{x}\right)$$

Let $\vec{x}_{\text{glob}}$ be a global minimum

$$f\left(\vec{x}_{\text{glob}}\right) < f\left(\vec{x}_{\text{loc}}\right)$$

Proof

Choose θ so that $\vec{x}_\theta := \theta \vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}$ satisfies

$$||\vec{x}_\theta - \vec{x}_{\text{loc}}||_2 \leq \gamma$$

then

$$f\left(\vec{x}_{\text{loc}}\right) \leq f\left(\vec{x}_\theta\right)$$

Proof

Choose θ so that $\vec{x}_{\theta} := \theta \vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}$ satisfies

$$||\vec{x}_{\theta} - \vec{x}_{\text{loc}}||_2 \leq \gamma$$

then

$$\begin{aligned} f\left(\vec{x}_{\text{loc}}\right) &\leq f\left(\vec{x}_{\theta}\right) \\ &= f\left(\theta \vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}\right) \end{aligned}$$

Proof

Choose θ so that $\vec{x}_\theta := \theta \vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}$ satisfies

$$||\vec{x}_\theta - \vec{x}_{\text{loc}}||_2 \leq \gamma$$

then

$$\begin{aligned} f\left(\vec{x}_{\text{loc}}\right) &\leq f\left(\vec{x}_\theta\right) \\ &= f\left(\theta \vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}\right) \\ &\leq \theta f\left(\vec{x}_{\text{loc}}\right) + (1-\theta)\, f\left(\vec{x}_{\text{glob}}\right) \quad \text{by convexity of } f \end{aligned}$$

Proof

Choose θ so that $\vec{x}_\theta := \theta \vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}$ satisfies

$$||\vec{x}_\theta - \vec{x}_{\text{loc}}||_2 \leq \gamma$$

then

$$\begin{aligned} f\left(\vec{x}_{\text{loc}}\right) &\leq f\left(\vec{x}_\theta\right) \\ &= f\left(\theta \vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}\right) \\ &\leq \theta f\left(\vec{x}_{\text{loc}}\right) + (1-\theta)\, f\left(\vec{x}_{\text{glob}}\right) \quad \text{by convexity of } f \\ &< f\left(\vec{x}_{\text{loc}}\right) \quad \text{because } f\left(\vec{x}_{\text{glob}}\right) < f\left(\vec{x}_{\text{loc}}\right) \end{aligned}$$

Norm

Let $\mathcal{V}$ be a vector space, a norm is a function $||\cdot||$ from $\mathcal{V}$ to $\mathbb{R}$ with the following properties

- It is homogeneous. For any scalar α and any $\vec{x} \in \mathcal{V}$

$$||\alpha \, \vec{x}|| = |\alpha| \, ||\vec{x}|| \, .$$

- It satisfies the triangle inequality

$$||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}|| \, .$$

  In particular, $||\vec{x}|| \geq 0$

- $||\vec{x}|| = 0$ implies $\vec{x} = \vec{0}$

Norms are convex

For any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and any $\theta \in (0,1)$

$$||\theta \vec{x} + (1-\theta)\,\vec{y}||$$

Norms are convex

For any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$||\theta \vec{x} + (1 - \theta)\, \vec{y}|| \leq ||\theta \vec{x}|| + ||(1 - \theta)\, \vec{y}||$$

Norms are convex

For any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$\begin{aligned} ||\theta\vec{x} + (1-\theta)\,\vec{y}|| &\leq ||\theta\vec{x}|| + ||(1-\theta)\,\vec{y}|| \\ &= \theta\,||\vec{x}|| + (1-\theta)\,||\vec{y}|| \end{aligned}$$

Composition of convex and affine function

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then for any $A \in \mathbb{R}^{n \times m}$ and $\vec{b} \in \mathbb{R}^n$

$$h(\vec{x}) := f\left(A\vec{x} + \vec{b}\right)$$

is convex

Consequence:

$$f(\vec{x}) := \left|\left|A\vec{x} + \vec{b}\right|\right|$$

is convex for any A and $\vec{b}$

Composition of convex and affine function

$$h\left(\theta\vec{x} + (1-\theta)\,\vec{y}\right)$$

Composition of convex and affine function

$$h\left(\theta\vec{x} + (1-\theta)\,\vec{y}\right) = f\left(\theta\left(A\vec{x} + \vec{b}\right) + (1-\theta)\left(A\vec{y} + \vec{b}\right)\right)$$

Composition of convex and affine function

$$\begin{aligned} h\left(\theta\vec{x} + (1-\theta)\,\vec{y}\right) &= f\left(\theta\left(A\vec{x} + \vec{b}\right) + (1-\theta)\left(A\vec{y} + \vec{b}\right)\right) \\ &\leq \theta f\left(A\vec{x} + \vec{b}\right) + (1-\theta)\, f\left(A\vec{y} + \vec{b}\right) \end{aligned}$$

Composition of convex and affine function

$$
\begin{aligned}
h\left(\theta\vec{x} + (1-\theta)\,\vec{y}\right) &= f\left(\theta\left(A\vec{x} + \vec{b}\right) + (1-\theta)\left(A\vec{y} + \vec{b}\right)\right) \\
&\leq \theta f\left(A\vec{x} + \vec{b}\right) + (1-\theta)\, f\left(A\vec{y} + \vec{b}\right) \\
&= \theta\, h\left(\vec{x}\right) + (1-\theta)\, h\left(\vec{y}\right)
\end{aligned}
$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$||2\vec{x}||_0$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$||2\vec{x}||_0 = ||\vec{x}||_0$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} ||2\vec{x}||_0 &= ||\vec{x}||_0 \\ &\neq 2\,||\vec{x}||_0 \end{aligned}$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} \|2\vec{x}\|_0 &= \|\vec{x}\|_0 \\ &\neq 2\,\|\vec{x}\|_0 \end{aligned}$$

Not convex

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} ||2\vec{x}||_0 &= ||\vec{x}||_0 \\ &\neq 2\,||\vec{x}||_0 \end{aligned}$$

Not convex

Let $\vec{x} := \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$ and $\vec{y} := \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, for any $\theta \in (0, 1)$

$$||\theta\vec{x} + (1-\theta)\,\vec{y}||_0$$

$$\theta\,||\vec{x}||_0 + (1-\theta)\,||\vec{y}||_0$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} ||2\vec{x}||_0 &= ||\vec{x}||_0 \\ &\neq 2\,||\vec{x}||_0 \end{aligned}$$

Not convex

Let $\vec{x} := \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$ and $\vec{y} := \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)$, for any $\theta \in (0,1)$

$$||\theta\vec{x} + (1-\theta)\,\vec{y}||_0 \;= 2$$

$$\theta\,||\vec{x}||_0 + (1-\theta)\,||\vec{y}||_0$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} ||2\vec{x}||_0 &= ||\vec{x}||_0 \\ &\neq 2\,||\vec{x}||_0 \end{aligned}$$

Not convex

Let $\vec{x} := \binom{1}{0}$ and $\vec{y} := \binom{0}{1}$, for any $\theta \in (0, 1)$

$$||\theta\vec{x} + (1-\theta)\,\vec{y}||_0 = 2$$

$$\theta\,||\vec{x}||_0 + (1-\theta)\,||\vec{y}||_0 = 1$$

Promoting sparsity

Finding sparse vectors consistent with data is often very useful

Toy problem: Find t such that

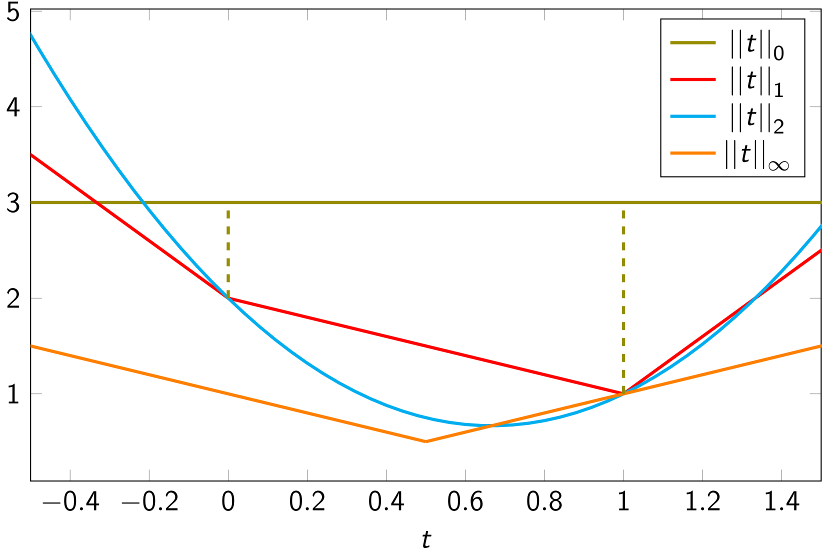
$$\vec{v}_t := \begin{bmatrix} t \\ t-1 \\ t-1 \end{bmatrix}$$

is sparse

Strategy: Minimize

$$f(t) := ||\vec{v}_t||$$

Promoting sparsity



The rank is not convex

The rank of matrices in $\mathbb{R}^{n\times n}$ interpreted as a function from $\mathbb{R}^{n\times n}$ to $\mathbb{R}$ is not convex

The rank is not convex

The rank of matrices in $\mathbb{R}^{n\times n}$ interpreted as a function from $\mathbb{R}^{n\times n}$ to $\mathbb{R}$ is not convex

$$X := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad Y := \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

For any $\theta \in (0,1)$

$$\operatorname{rank}\left(\theta X + (1-\theta)\, Y\right)$$

$$\theta \operatorname{rank}(X) + (1-\theta) \operatorname{rank}(Y)$$

The rank is not convex

The rank of matrices in $\mathbb{R}^{n\times n}$ interpreted as a function from $\mathbb{R}^{n\times n}$ to $\mathbb{R}$ is not convex

$$X := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad Y := \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

For any $\theta \in (0,1)$

$$\operatorname{rank}\left(\theta X + (1-\theta)\, Y\right) = 2$$

$$\theta \operatorname{rank}(X) + (1-\theta) \operatorname{rank}(Y)$$

The rank is not convex

The rank of matrices in $\mathbb{R}^{n\times n}$ interpreted as a function from $\mathbb{R}^{n\times n}$ to $\mathbb{R}$ is not convex

$$X := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad Y := \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

For any $\theta \in (0, 1)$

$$\operatorname{rank}\left(\theta X + (1-\theta)\, Y\right) = 2$$

$$\theta \operatorname{rank}(X) + (1-\theta) \operatorname{rank}(Y) = 1$$

Matrix norms

Frobenius norm

$$||A||_{\mathsf{F}} := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}$$

Operator norm

$$||A|| := \max_{\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\}} ||A\,\vec{x}||_2 = \sigma_1$$

Nuclear norm

$$||A||_* := \sum_{i=1}^{\min\{m,n\}} \sigma_i$$

Promoting low-rank structure

Finding low-rank matrices consistent with data is often very useful

Toy problem: Find t such that

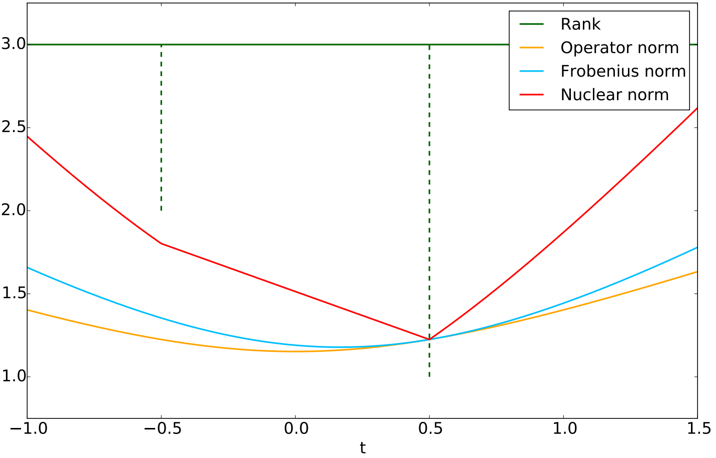
$$M(t) := \begin{bmatrix} 0.5 + t & 1 & 1 \\ 0.5 & 0.5 & t \\ 0.5 & 1 - t & 0.5 \end{bmatrix},$$

is low rank

Strategy: Minimize

$$f(t) := ||M(t)||$$

Promoting low-rank structure



Convexity

Differentiable convex functions

Minimizing differentiable convex functions

Gradient

$$\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial \vec{x}[1]} \\ \frac{\partial f(\vec{x})}{\partial \vec{x}[2]} \\ \cdots \\ \frac{\partial f(\vec{x})}{\partial \vec{x}[n]} \end{bmatrix}$$

If the gradient exists at every point, the function is said to be differentiable

Directional derivative

Encodes first-order rate of change in a particular direction

$$f'_{\vec{u}}(\vec{x}) := \lim_{h \to 0} \frac{f(\vec{x} + h\vec{u}) - f(\vec{x})}{h}$$
$$= \langle \nabla f(\vec{x}), \vec{u} \rangle$$

where $||u||_2 = 1$

Direction of maximum variation

∇f is direction of maximum increase

-∇f is direction of maximum decrease

$$\left| f'_{\vec{u}}(\vec{x}) \right| = \left| \nabla f(\vec{x})^T \vec{u} \right|$$

Direction of maximum variation

∇f is direction of maximum increase

-∇f is direction of maximum decrease

$$\begin{aligned}\left|f'_{\vec{u}}(\vec{x})\right| &= \left|\nabla f(\vec{x})^T \vec{u}\right| \\ &\leq \|\nabla f(\vec{x})\|_2 \|\vec{u}\|_2 \qquad \text{Cauchy-Schwarz inequality}\end{aligned}$$

Direction of maximum variation

∇f is direction of maximum increase

-∇f is direction of maximum decrease

$$
\begin{aligned}
\left| f'_{\vec{u}}(\vec{x}) \right| &= \left| \nabla f(\vec{x})^T \vec{u} \right| \\
&\leq ||\nabla f(\vec{x})||_2 \, ||\vec{u}||_2 \qquad \text{Cauchy-Schwarz inequality} \\
&= ||\nabla f(\vec{x})||_2
\end{aligned}
$$

Direction of maximum variation

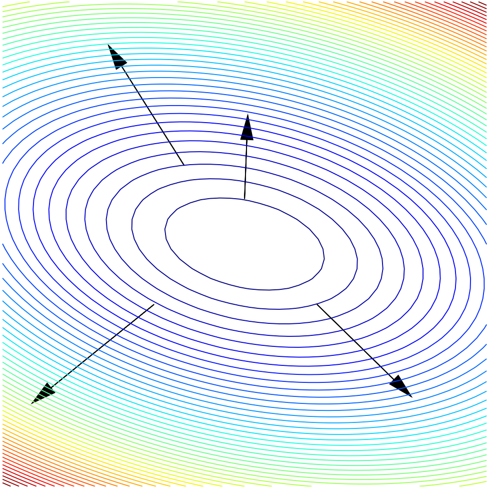
∇f is direction of maximum increase

-∇f is direction of maximum decrease

$$\begin{aligned} \left| f'_{\vec{u}}(\vec{x}) \right| &= \left| \nabla f \left(\vec{x} \right)^T \vec{u} \right| \\ &\leq \left|\left| \nabla f \left(\vec{x} \right) \right|\right|_2 \left|\left| \vec{u} \right|\right|_2 \qquad \text{Cauchy-Schwarz inequality} \\ &= \left|\left| \nabla f \left(\vec{x} \right) \right|\right|_2 \end{aligned}$$

equality holds if and only if $\vec{u} = \pm \frac{\nabla f(\vec{x})}{||\nabla f(\vec{x})||_2}$

Gradient



First-order approximation

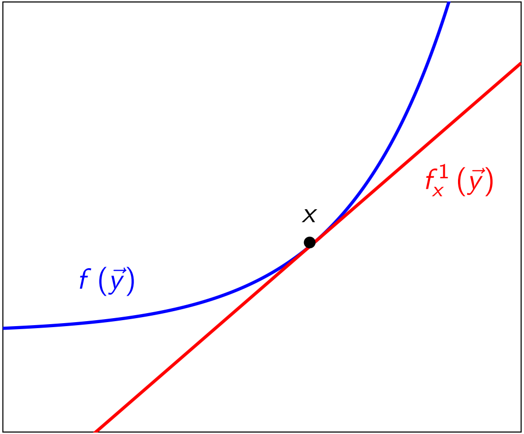
The first-order or linear approximation of $f : \mathbb{R}^n \to \mathbb{R}$ at $\vec{x}$ is

$$f^1_{\vec{x}}(\vec{y}) := f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x})$$

If f is continuously differentiable at $\vec{x}$

$$\lim_{\vec{y} \to \vec{x}} \frac{f(\vec{y}) - f^1_{\vec{x}}(\vec{y})}{||\vec{y} - \vec{x}||_2} = 0$$

First-order approximation



Convexity

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for every $\vec{x}, \vec{y} \in \mathbb{R}^n$

$$f(\vec{y}) \geq f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x})$$

It is strictly convex if and only if

$$f(\vec{y}) > f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x})$$

Optimality condition

If f is convex and $\nabla f(\vec{x}) = 0$, then for any $\vec{y} \in \mathbb{R}$

$$f(\vec{y}) \geq f(\vec{x})$$

If f is strictly convex then for any $\vec{y} \neq \vec{x}$

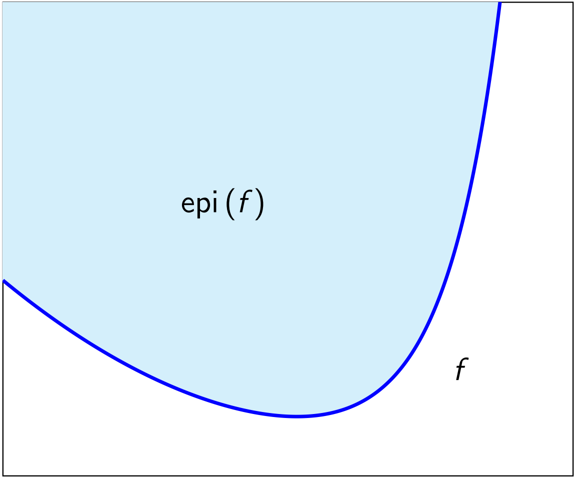
$$f(\vec{y}) > f(\vec{x})$$

Epigraph

The epigraph of $f : \mathbb{R}^n \to \mathbb{R}$ is

$$\operatorname{epi}(f) := \left\{ \vec{x} \mid f\left(\begin{bmatrix} \vec{x}[1] \\ \cdots \\ \vec{x}[n] \end{bmatrix} \right) \leq \vec{x}[n+1] \right\}$$

Epigraph



Supporting hyperplane

A hyperplane $\mathcal{H}$ is a supporting hyperplane of a set $\mathcal{S}$ at $\vec{x}$ if

- $\mathcal{H}$ and $\mathcal{S}$ intersect at $\vec{x}$
- $\mathcal{S}$ is contained in one of the half-spaces bounded by $\mathcal{H}$

Geometric intuition

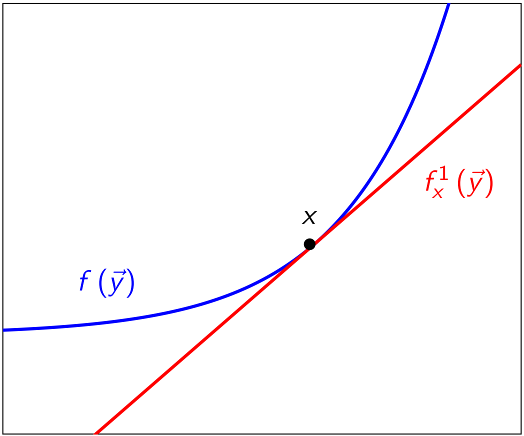
Geometrically, f is convex if and only if for every $\vec{x}$ the plane

$$\mathcal{H}_{f,\vec{x}} := \left\{ \vec{y} \mid \vec{y}[n+1] = f^1_{\vec{x}} \left(\begin{bmatrix} \vec{y}[1] \\ \cdots \\ \vec{y}[n] \end{bmatrix} \right) \right\}$$

is a supporting hyperplane of the epigraph at $\vec{x}$

If $\nabla f(\vec{x}) = 0$ the hyperplane is horizontal

Convexity



Hessian matrix

If f has a Hessian matrix at every point, it is twice differentiable

$$\nabla^2 f(\vec{x}) = \begin{bmatrix} \frac{\partial^2 f(\vec{x})}{\partial \vec{x}[1]^2} & \frac{\partial^2 f(\vec{x})}{\partial \vec{x}[1]\partial \vec{x}[2]} & \cdots & \frac{\partial^2 f(\vec{x})}{\partial \vec{x}[1]\partial \vec{x}[n]} \\ \frac{\partial^2 f(\vec{x})}{\partial \vec{x}[1]\partial \vec{x}[2]} & \frac{\partial^2 f(\vec{x})}{\partial \vec{x}[1]^2} & \cdots & \frac{\partial^2 f(\vec{x})}{\partial \vec{x}[2]\partial \vec{x}[n]} \\ & & \cdots & \\ \frac{\partial^2 f(\vec{x})}{\partial \vec{x}[1]\partial \vec{x}[n]} & \frac{\partial^2 f(\vec{x})}{\partial \vec{x}[2]\partial \vec{x}[n]} & \cdots & \frac{\partial^2 f(\vec{x})}{\partial \vec{x}[n]^2} \end{bmatrix}$$

Curvature

The second directional derivative $f''_{\vec{u}}$ of f at $\vec{x}$ equals

$$f''_{\vec{u}}\left(\vec{x}\right) = \vec{u}^{\,T} \nabla^2 f\left(\vec{x}\right) \vec{u}$$

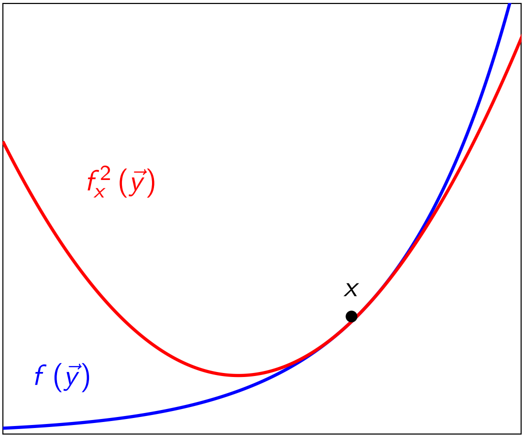
for any unit-norm vector $\vec{u} \in \mathbb{R}^n$

Second-order approximation

The second-order or quadratic approximation of f at $\vec{x}$ is

$$f_{\vec{x}}^{2}\left(\vec{y}\right) := f\left(\vec{x}\right) + \nabla f\left(\vec{x}\right)\left(\vec{y} - \vec{x}\right) + \frac{1}{2}\left(\vec{y} - \vec{x}\right)^{T} \nabla^{2} f\left(\vec{x}\right)\left(\vec{y} - \vec{x}\right)$$

Second-order approximation



Quadratic form

Second order polynomial in several dimensions

$$q(\vec{x}) := \vec{x}^T A \vec{x} + \vec{b}^T \vec{x} + c$$

parametrized by symmetric matrix $A \in \mathbb{R}^{n \times n}$, a vector $\vec{b} \in \mathbb{R}^n$ and a constant c

Quadratic approximation

The quadratic approximation $f^2_{\vec{x}} : \mathbb{R}^n \to \mathbb{R}$ at $\vec{x} \in \mathbb{R}^n$ of a twice-continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies

$$\lim_{\vec{y} \to \vec{x}} \frac{f(\vec{y}) - f^2_{\vec{x}}(\vec{y})}{||\vec{y} - \vec{x}||_2^2} = 0$$

Eigendecomposition of symmetric matrices

Let $A = U \Lambda U^T$ be the eigendecomposition of a symmetric matrix A

Eigenvalues: $\lambda_1 \geq \cdots \geq \lambda_n$ (which can be negative or 0)

Eigenvectors: $\vec{u}_1, \ldots, \vec{u}_n$, orthonormal basis

$$\lambda_1 = \max_{\{||\vec{x}||_2 = 1 \mid \vec{x} \in \mathbb{R}^n\}} \vec{x}^T A \vec{x}$$

$$\vec{u}_1 = \arg\max_{\{||\vec{x}||_2 = 1 \mid \vec{x} \in \mathbb{R}^n\}} \vec{x}^T A \vec{x}$$

$$\lambda_n = \min_{\{||\vec{x}||_2 = 1 \mid \vec{x} \in \mathbb{R}^n\}} \vec{x}^T A \vec{x}$$

$$\vec{u}_n = \arg\min_{\{||\vec{x}||_2 = 1 \mid \vec{x} \in \mathbb{R}^n\}} \vec{x}^T A \vec{x}$$

Maximum and minimum curvature

Let $\nabla^2 f(\vec{x}) = U \Lambda U^T$ be the eigendecomposition of the Hessian at $\vec{x}$

Direction of maximum curvature: $\vec{u}_1$

Direction of minimum curvature (or maximum negative curvature): $\vec{u}_n$

Positive semidefinite matrices

For any $\vec{x}$

$$\vec{x}^T A \vec{x} = \vec{x}^T U \Lambda U^T \vec{x}$$
$$= \sum_{i=1}^{n} \lambda_i \langle \vec{u}_i, \vec{x} \rangle^2$$

All eigenvalues are nonnegative if and only if

$$\vec{x}^T A \vec{x} \geq 0$$

for all $\vec{x}$

The matrix is positive semidefinite

Positive (negative) (semi)definite matrices

Positive (semi)definite: all eigenvalues are positive (nonnegative), equivalently for all $\vec{x}$

$$\vec{x}^T A\vec{x} > (\geq)\ 0$$

Quadratic form: All directions have positive curvature

Negative (semi)definite: all eigenvalues are negative (nonpositive), equivalently for all $\vec{x}$

$$\vec{x}^T A\vec{x} < (\leq)\ 0$$

Quadratic form: All directions have negative curvature

Convexity

A twice-differentiable function $g : \mathbb{R} \to \mathbb{R}$ is convex if and only if

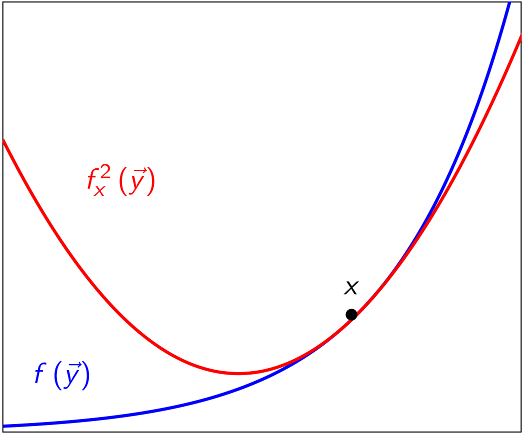
$$g''(x) \geq 0$$

for all $x \in \mathbb{R}$

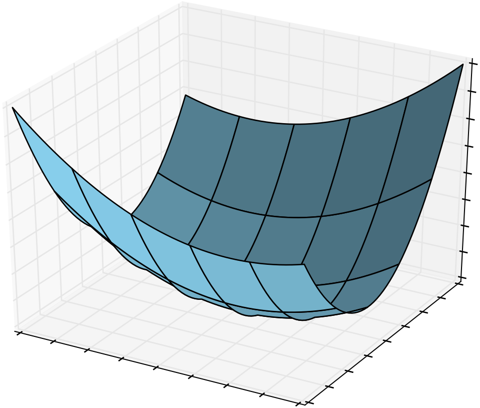
A twice-differentiable function in $\mathbb{R}^n$ is convex if and only if their Hessian is positive semidefinite at every point

If the Hessian is positive definite, the function is strictly convex

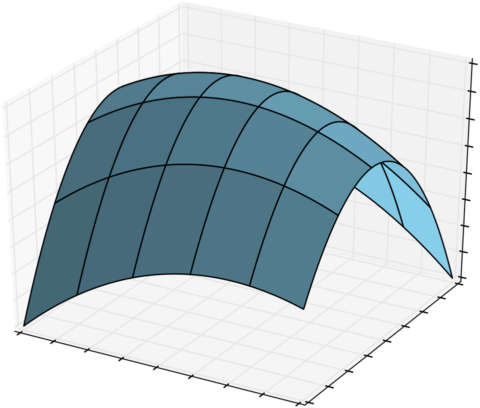
Second-order approximation



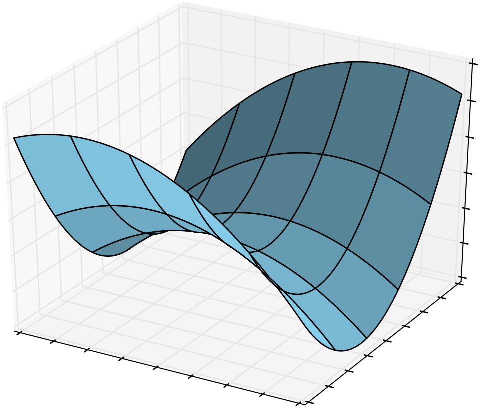
Convex



Concave



Neither



Convexity

Differentiable convex functions

Minimizing differentiable convex functions

Problem

Challenge: Minimizing differentiable convex functions

$$\min_{\vec{x} \in \mathbb{R}^n} \quad f(\vec{x})$$

Gradient descent

Intuition: Make local progress in the steepest direction $-\nabla f(\vec{x})$

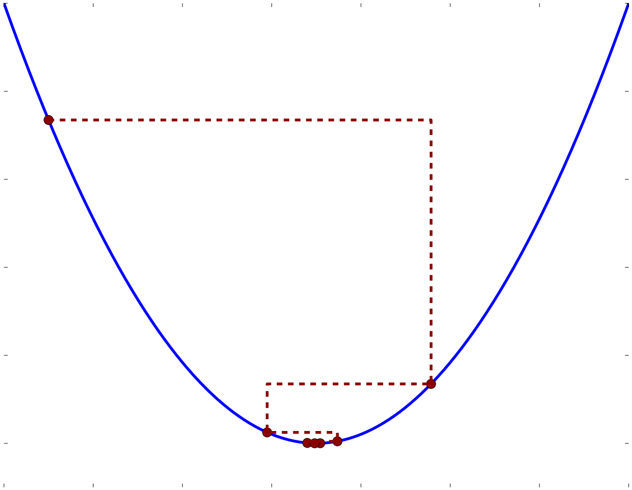
Set the initial point $\vec{x}^{(0)}$ to an arbitrary value

Update by setting

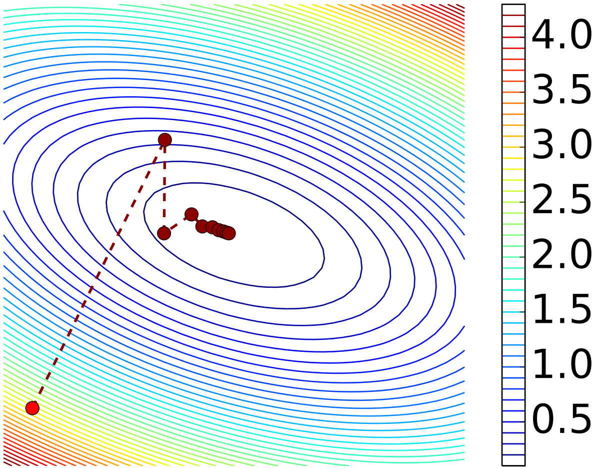
$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

where $\alpha_k > 0$ is the step size, until a stopping criterion is met

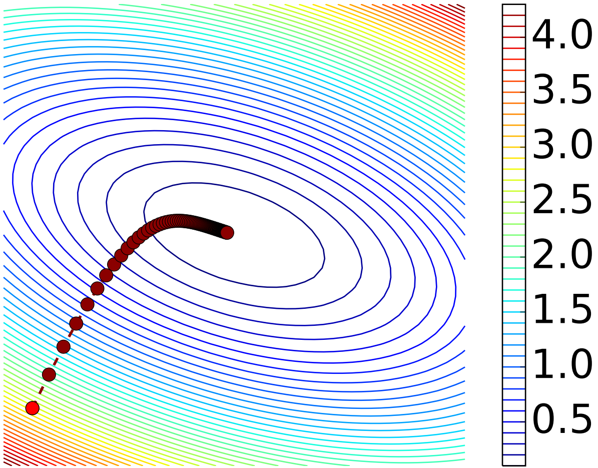
Gradient descent



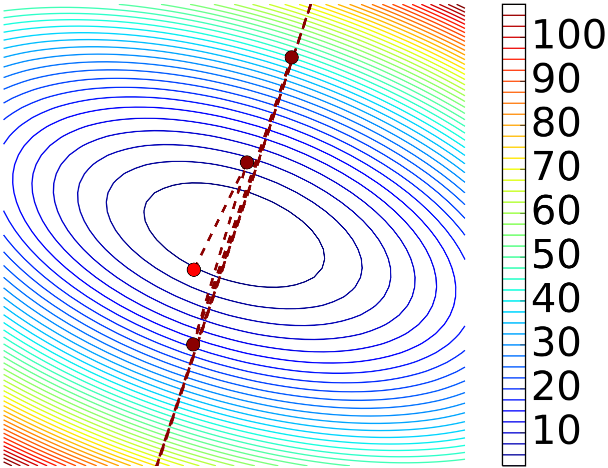
Gradient descent



Small step size



Large step size



Line search

Idea: Find minimum of

$$\begin{aligned} \alpha_k &:= \arg\min_{\alpha} h(\alpha) \\ &= \arg\min_{\alpha \in \mathbb{R}} f\left(\vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right) \end{aligned}$$

Backtracking line search with Armijo rule

Given $\alpha^0 \geq 0$ and $\beta, \eta \in (0,1)$, set $\alpha_k := \alpha^0 \beta^i$ for smallest i such that

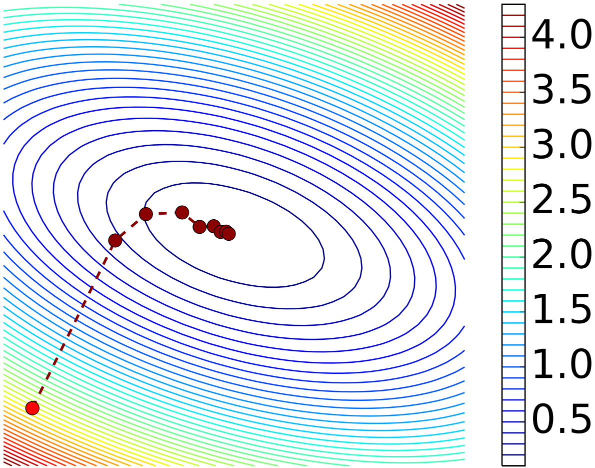
$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

satisfies

$$f\left(\vec{x}^{(k+1)}\right) \leq f\left(\vec{x}^{(k)}\right) - \frac{1}{2}\alpha_k \left|\left|\nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2$$

a condition known as Armijo rule

Backtracking line search with Armijo rule



Gradient descent for least squares

Aim: Use n examples

$$\left(y^{(1)}, \vec{x}^{(1)}\right), \left(y^{(2)}, \vec{x}^{(2)}\right), \ldots, \left(y^{(n)}, \vec{x}^{(n)}\right)$$

to fit a linear model by minimizing least-squares cost function

$$\text{minimize}_{\vec{\beta} \in \mathbb{R}^p} \left|\left|\vec{y} - X\vec{\beta}\right|\right|_2^2$$

Gradient descent for least squares

The gradient of the quadratic function

$$f(\vec{\beta}) := \left|\left|\vec{y} - X\vec{\beta}\right|\right|_2^2$$
$$= \vec{\beta}^T X^T X \vec{\beta} - 2\vec{\beta}^T X^T \vec{y} + \vec{y}^T \vec{y}$$

equals

$$\nabla f(\vec{\beta})$$

Gradient descent for least squares

The gradient of the quadratic function

$$
\begin{aligned}
f(\vec{\beta}) &:= \left|\left|\vec{y} - X\vec{\beta}\right|\right|_2^2 \\
&= \vec{\beta}^T X^T X \vec{\beta} - 2\vec{\beta}^T X^T \vec{y} + \vec{y}^T \vec{y}
\end{aligned}
$$

equals

$$
\nabla f(\vec{\beta}) = 2X^T X \vec{\beta} - 2X^T \vec{y}
$$

Gradient descent for least squares

The gradient of the quadratic function

$$f(\vec{\beta}) := \left|\left|\vec{y} - X\vec{\beta}\right|\right|_2^2$$
$$= \vec{\beta}^T X^T X \vec{\beta} - 2\vec{\beta}^T X^T \vec{y} + \vec{y}^T \vec{y}$$

equals

$$\nabla f(\vec{\beta}) = 2X^T X\vec{\beta} - 2X^T \vec{y}$$

Gradient descent updates are

$$\vec{\beta}^{(k+1)} = \vec{\beta}^{(k)} + 2\alpha_k X^T \left(\vec{y} - X\vec{\beta}^{(k)}\right)$$

Gradient descent for least squares

The gradient of the quadratic function

$$\begin{aligned} f(\vec{\beta}) &:= \left|\left|\vec{y} - X\vec{\beta}\right|\right|_2^2 \\ &= \vec{\beta}^T X^T X \vec{\beta} - 2\vec{\beta}^T X^T \vec{y} + \vec{y}^T \vec{y} \end{aligned}$$

equals

$$\nabla f(\vec{\beta}) = 2X^T X \vec{\beta} - 2X^T \vec{y}$$

Gradient descent updates are

$$\begin{aligned} \vec{\beta}^{(k+1)} &= \vec{\beta}^{(k)} + 2\alpha_k X^T \left(\vec{y} - X\vec{\beta}^{(k)}\right) \\ &= \vec{\beta}^{(k)} + 2\alpha_k \sum_{i=1}^{n} \left(\vec{y}^{(i)} - \langle x^{(i)}, \vec{\beta}^{(k)} \rangle\right) x^{(i)} \end{aligned}$$

Gradient ascent for logistic regression

Aim: Use n examples

$$\left(y^{(1)}, \vec{x}^{(1)}\right), \left(y^{(2)}, \vec{x}^{(2)}\right), \ldots, \left(y^{(n)}, \vec{x}^{(n)}\right)$$

to fit logistic-regression model by maximizing log-likelihood cost function

$$f(\vec{\beta}) := \sum_{i=1}^{n} y^{(i)} \log g\left(\langle \vec{x}^{(i)}, \vec{\beta} \rangle\right) + \left(1 - y^{(i)}\right) \log\left(1 - g\left(\langle \vec{x}^{(i)}, \vec{\beta} \rangle\right)\right)$$

where

$$g(t) = \frac{1}{1 - \exp -t}$$

Gradient ascent for logistic regression

$$g'(t) = g(t)(1 - g(t))$$
$$(1 - g(t))' = -g(t)(1 - g(t))$$

The gradient of the cost function equals

$\nabla f(\vec{\beta})$

Gradient ascent for logistic regression

$$g'(t) = g(t)(1 - g(t))$$
$$(1 - g(t))' = -g(t)(1 - g(t))$$

The gradient of the cost function equals

$$\nabla f(\vec{\beta}) = \sum_{i=1}^{n} y^{(i)} \left(1 - g(\langle \vec{x}^{(i)}, \vec{\beta} \rangle)\right) \vec{x}^{(i)} - \left(1 - y^{(i)}\right) g(\langle \vec{x}^{(i)}, \vec{\beta} \rangle) \vec{x}^{(i)}$$

Gradient ascent for logistic regression

$$g'(t) = g(t)(1 - g(t))$$
$$(1 - g(t))' = -g(t)(1 - g(t))$$

The gradient of the cost function equals

$$\nabla f(\vec{\beta}) = \sum_{i=1}^{n} y^{(i)} \left(1 - g(\langle \vec{x}^{(i)}, \vec{\beta} \rangle)\right) \vec{x}^{(i)} - \left(1 - y^{(i)}\right) g(\langle \vec{x}^{(i)}, \vec{\beta} \rangle) \vec{x}^{(i)}$$

The gradient ascent updates are

$\vec{\beta}^{(k+1)}$

Gradient ascent for logistic regression

$$g'(t) = g(t)(1 - g(t))$$
$$(1 - g(t))' = -g(t)(1 - g(t))$$

The gradient of the cost function equals

$$\nabla f(\vec{\beta}) = \sum_{i=1}^{n} y^{(i)} \left(1 - g(\langle \vec{x}^{(i)}, \vec{\beta} \rangle)\right) \vec{x}^{(i)} - \left(1 - y^{(i)}\right) g(\langle \vec{x}^{(i)}, \vec{\beta} \rangle) \vec{x}^{(i)}$$

The gradient ascent updates are

$$\vec{\beta}^{(k+1)} := \vec{\beta}^{(k)} + \alpha_k \sum_{i=1}^{n} y^{(i)} \left(1 - g(\langle \vec{x}^{(i)}, \vec{\beta}^{(k)} \rangle)\right) \vec{x}^{(i)} - \left(1 - y^{(i)}\right) g(\langle \vec{x}^{(i)}, \vec{\beta}^{(k)} \rangle) \vec{x}^{(i)}$$

Convergence of gradient descent

Does the method converge?

How fast (slow)?

For what step sizes?

Convergence of gradient descent

Does the method converge?

How fast (slow)?

For what step sizes?

Depends on function

Lipschitz continuity

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous if for any $\vec{x}, \vec{y} \in \mathbb{R}^n$

$$||f(\vec{y}) - f(\vec{x})||_2 \leq L \, ||\vec{y} - \vec{x}||_2 \, .$$

L is the Lipschitz constant

Lipschitz-continuous gradients

If ∇f is Lipschitz continuous with Lipschitz constant L

$$\left|\left|\nabla f\left(\vec{y}\right)-\nabla f\left(\vec{x}\right)\right|\right|_2 \leq L\left|\left|\vec{y}-\vec{x}\right|\right|_2$$

then for any $\vec{x}, \vec{y} \in \mathbb{R}^n$ we have a quadratic upper bound

$$f\left(\vec{y}\right) \leq f\left(\vec{x}\right)+\nabla f\left(\vec{x}\right)^T\left(\vec{y}-\vec{x}\right)+\frac{L}{2}\left|\left|\vec{y}-\vec{x}\right|\right|_2^2$$

Local progress of gradient descent

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

$$f\left(\vec{x}^{(k+1)}\right)$$

Local progress of gradient descent

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

$$\begin{aligned} & f\left(\vec{x}^{(k+1)}\right) \\ & \leq f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x}^{(k+1)} - \vec{x}^{(k)}\right) + \frac{L}{2}\left|\left|\vec{x}^{(k+1)} - \vec{x}^{(k)}\right|\right|_2^2 \end{aligned}$$

Local progress of gradient descent

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

$$\begin{aligned}
& f\left(\vec{x}^{(k+1)}\right) \\
& \leq f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x}^{(k+1)} - \vec{x}^{(k)}\right) + \frac{L}{2} \left|\left|\vec{x}^{(k+1)} - \vec{x}^{(k)}\right|\right|_2^2 \\
& = f\left(\vec{x}^{(k)}\right) - \alpha_k \left(1 - \frac{\alpha_k L}{2}\right) \left|\left|\nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2
\end{aligned}$$

Local progress of gradient descent

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

$$\begin{aligned}
&f\left(\vec{x}^{(k+1)}\right) \\
&\leq f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T\left(\vec{x}^{(k+1)} - \vec{x}^{(k)}\right) + \frac{L}{2}\left|\left|\vec{x}^{(k+1)} - \vec{x}^{(k)}\right|\right|_2^2 \\
&= f\left(\vec{x}^{(k)}\right) - \alpha_k\left(1 - \frac{\alpha_k L}{2}\right)\left|\left|\nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2
\end{aligned}$$

If $\alpha_k \leq \frac{1}{L}$

$$f\left(\vec{x}^{(k+1)}\right) \leq f\left(\vec{x}^{(k)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2$$

Convergence of gradient descent

- f is convex
- ∇f is L-Lipschitz continuous
- There exists a point $\vec{x}^*$ at which f achieves a finite minimum
- The step size is set to $\alpha_k := \alpha \leq 1/L$

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^*\right) \leq \frac{\left|\left|\vec{x}^{(0)} - \vec{x}^*\right|\right|_2^2}{2\,\alpha\,k}$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) \leq f\left(\vec{x}^{(k-1)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

$$f\left(\vec{x}^{(k-1)}\right) + \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^* - \vec{x}^{(k-1)}\right) \leq f\left(\vec{x}^*\right)$$

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^*\right)$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) \leq f\left(\vec{x}^{(k-1)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

$$f\left(\vec{x}^{(k-1)}\right) + \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^* - \vec{x}^{(k-1)}\right) \leq f\left(\vec{x}^*\right)$$

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^*\right)$$

$$\leq f\left(\vec{x}^{(k-1)}\right) - f\left(\vec{x}^*\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) \leq f\left(\vec{x}^{(k-1)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

$$f\left(\vec{x}^{(k-1)}\right) + \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^* - \vec{x}^{(k-1)}\right) \leq f\left(\vec{x}^*\right)$$

$$\begin{aligned} & f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^*\right) \\ & \leq f\left(\vec{x}^{(k-1)}\right) - f\left(\vec{x}^*\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \\ & \leq \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{(k-1)} - \vec{x}^*\right) - \frac{\alpha}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \end{aligned}$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) \leq f\left(\vec{x}^{(k-1)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

$$f\left(\vec{x}^{(k-1)}\right) + \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{*} - \vec{x}^{(k-1)}\right) \leq f\left(\vec{x}^{*}\right)$$

$$\begin{aligned}
&f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \\
&\leq f\left(\vec{x}^{(k-1)}\right) - f\left(\vec{x}^{*}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \\
&\leq \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{(k-1)} - \vec{x}^{*}\right) - \frac{\alpha}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \\
&= \frac{1}{2\,\alpha}\left(\left|\left|\vec{x}^{(k-1)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k-1)} - \vec{x}^{*} - \alpha\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2\right)
\end{aligned}$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) \leq f\left(\vec{x}^{(k-1)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

$$f\left(\vec{x}^{(k-1)}\right) + \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^* - \vec{x}^{(k-1)}\right) \leq f\left(\vec{x}^*\right)$$

$$\begin{aligned}
& f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^*\right) \\
& \leq f\left(\vec{x}^{(k-1)}\right) - f\left(\vec{x}^*\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \\
& \leq \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{(k-1)} - \vec{x}^*\right) - \frac{\alpha}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \\
& = \frac{1}{2\,\alpha}\left(\left|\left|\vec{x}^{(k-1)} - \vec{x}^*\right|\right|_2^2 - \left|\left|\vec{x}^{(k-1)} - \vec{x}^* - \alpha\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2\right) \\
& = \frac{1}{2\,\alpha}\left(\left|\left|\vec{x}^{(k-1)} - \vec{x}^*\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^*\right|\right|_2\right)
\end{aligned}$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right)$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \leq \frac{1}{k}\sum_{i=1}^{k} f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right)$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \leq \frac{1}{k}\sum_{i=1}^{k} f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \qquad \text{never increases}$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \leq \frac{1}{k}\sum_{i=1}^{k} f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \qquad \text{never increases}$$

$$= \frac{1}{2\,\alpha\,k}\sum_{i=1}^{k} \left|\left|\vec{x}^{(k-1)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2$$

Convergence of gradient descent

$$
\begin{aligned}
f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) &\leq \frac{1}{k}\sum_{i=1}^{k} f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \qquad \text{never increases} \\
&= \frac{1}{2\,\alpha\,k}\sum_{i=1}^{k} \left|\left|\vec{x}^{(k-1)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2 \\
&= \frac{1}{2\,\alpha\,k}\left(\left|\left|\vec{x}^{(0)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2^2\right)
\end{aligned}
$$

Convergence of gradient descent

$$
\begin{aligned}
f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) &\leq \frac{1}{k}\sum_{i=1}^{k} f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \qquad \text{never increases} \\
&= \frac{1}{2\,\alpha\,k}\sum_{i=1}^{k} \left|\left|\vec{x}^{(k-1)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2 \\
&= \frac{1}{2\,\alpha\,k}\left(\left|\left|\vec{x}^{(0)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2^2\right) \\
&\leq \frac{\left|\left|\vec{x}^{(0)} - \vec{x}^{*}\right|\right|_2^2}{2\,\alpha\,k}
\end{aligned}
$$

Accelerated gradient descent

- Gradient descent takes $\mathcal{O}\left(1/\epsilon\right)$ to achieve an error of ϵ
- The optimal rate is $\mathcal{O}\left(1/\sqrt{\epsilon}\right)$
- Gradient descent can be accelerated by adding a momentum term

Accelerated gradient descent

Set the initial point $\vec{x}^{(0)}$ to an arbitrary value

Update by setting

$$y^{(k+1)} = x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)$$

$$x^{(k+1)} = \beta_k \, y^{(k+1)} + \gamma_k \, y^{(k)}$$

where α_k is the step size and $\beta_k > 0$ and $\gamma_k > 0$ are parameters

Digit classification

MNIST data

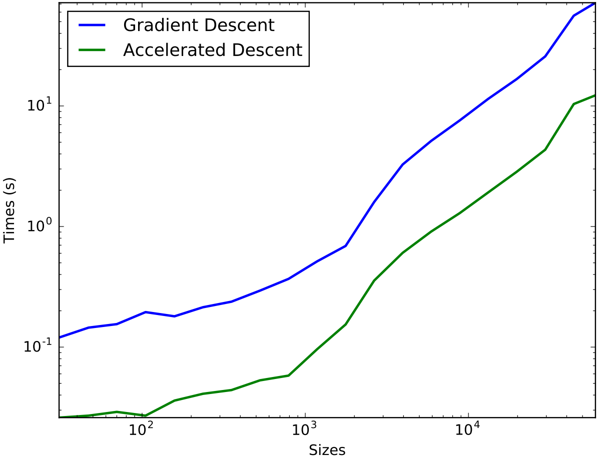
Aim: Determine whether a digit is a 5 or not

$\vec{x}_i$ is an image

$\vec{y}_i = 1$ or $\vec{y}_i = 0$ if image i is a 5 or not, respectively

We fit a logistic-regression model

Digit classification



Stochastic gradient descent

Cost functions to fit models are often additive

$$f\left(\vec{x}\right) = \frac{1}{m}\sum_{i=1}^{m} f_i\left(\vec{x}\right).$$

- Linear regression

$$\sum_{i=1}^{n}\left(y^{(i)} - \vec{x}^{(i)\,T}\vec{\beta}\right)^2 = \left|\left|\vec{y} - X\vec{\beta}\right|\right|_2^2$$

- Logistic regression

$$\sum_{i=1}^{n} y^{(i)} \log g\left(\langle \vec{x}^{(i)}, \vec{\beta}\rangle\right) + \left(1 - y^{(i)}\right)\log\left(1 - g\left(\langle \vec{x}^{(i)}, \vec{\beta}\rangle\right)\right)$$

Stochastic gradient descent

In *big data* regime (very large n), gradient descent is too slow

In some cases, data is acquired sequentially (online setting)

Stochastic gradient descent: update solution using a subset of the data

Stochastic gradient descent

Set the initial point $\vec{x}^{(0)}$ to an arbitrary value

Update by

1. Choosing a random subset of b indices $\boldsymbol{\mathcal{B}}$ ($b \ll m$ is the batch size)
2. Setting

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k m \sum_{i \in \boldsymbol{\mathcal{B}}} \nabla f_i \left(\vec{x}^{(k)} \right)$$

where α_k is the step size

Stochastic gradient descent

We replace ∇f by

$$\sum_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k)} \right) = \sum_{i=1}^{m} 1_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k+1)} \right)$$

Noisy estimate of ∇f

Unbiased if every example is in the batch with probability p

$$\mathrm{E} \left(\sum_{i=1}^{m} 1_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k)} \right) \right)$$

Stochastic gradient descent

We replace ∇f by

$$\sum_{i \in \mathcal{B}} \nabla f_i\left(\vec{x}^{(k)}\right) = \sum_{i=1}^{m} 1_{i \in \mathcal{B}} \nabla f_i\left(\vec{x}^{(k+1)}\right)$$

Noisy estimate of ∇f

Unbiased if every example is in the batch with probability p

$$\mathrm{E}\left(\sum_{i=1}^{m} 1_{i \in \mathcal{B}} \nabla f_i\left(\vec{x}^{(k)}\right)\right) = \sum_{i=1}^{m} \mathrm{E}\left(1_{i \in \mathcal{B}}\right) \nabla f_i\left(\vec{x}^{(k)}\right)$$

Stochastic gradient descent

We replace ∇f by

$$\sum_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k)}\right) = \sum_{i=1}^{m} 1_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k+1)}\right)$$

Noisy estimate of ∇f

Unbiased if every example is in the batch with probability p

$$\begin{aligned} \mathrm{E}\left(\sum_{i=1}^{m} 1_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k)}\right)\right) &= \sum_{i=1}^{m} \mathrm{E}\left(1_{i \in \mathcal{B}}\right) \nabla f_i \left(\vec{x}^{(k)}\right) \\ &= \sum_{i=1}^{m} \mathrm{P}\left(i \in \mathcal{B}\right) \nabla f_i \left(\vec{x}^{(k)}\right) \end{aligned}$$

Stochastic gradient descent

We replace ∇f by

$$\sum_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k)} \right) = \sum_{i=1}^{m} 1_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k+1)} \right)$$

Noisy estimate of ∇f

Unbiased if every example is in the batch with probability p

$$\begin{aligned} \mathrm{E}\left(\sum_{i=1}^{m} 1_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k)} \right) \right) &= \sum_{i=1}^{m} \mathrm{E}\left(1_{i \in \mathcal{B}} \right) \nabla f_i \left(\vec{x}^{(k)} \right) \\ &= \sum_{i=1}^{m} \mathrm{P}\left(i \in \mathcal{B} \right) \nabla f_i \left(\vec{x}^{(k)} \right) \\ &= p \nabla f \left(\vec{x}^{(k)} \right) \end{aligned}$$

Stochastic gradient descent

- Linear regression

$$\vec{\beta}^{(k+1)} := \vec{\beta}^{(k)} + 2\alpha_k \sum_{i \in \mathcal{B}} \left(\vec{y}^{(i)} - \langle x^{(i)}, \vec{\beta}^{(k)} \rangle \right) x^{(i)}$$

- Logistic regression

$$\vec{\beta}^{(k+1)} := \vec{\beta}^{(k)}$$
$$+ \alpha_k \sum_{i \in \mathcal{B}} y^{(i)} \left(1 - g(\langle \vec{x}^{(i)}, \vec{\beta}^{(k)} \rangle) \right) \vec{x}^{(i)} - \left(1 - y^{(i)} \right) g(\langle \vec{x}^{(i)}, \vec{\beta}^{(k)} \rangle) \vec{x}^{(i)}$$

Digit classification

MNIST data

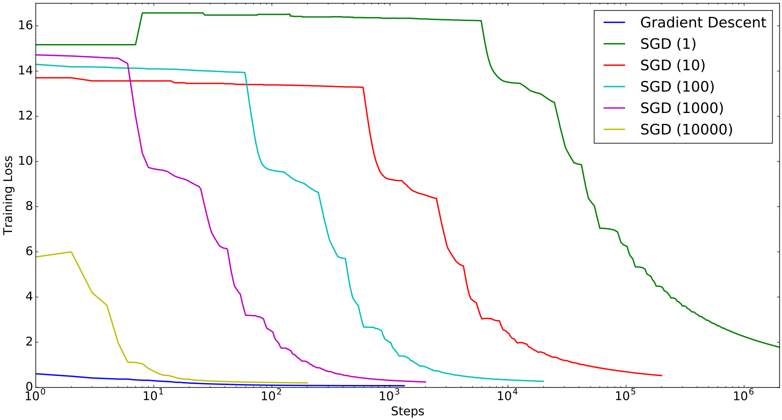
Aim: Determine whether a digit is a 5 or not

$\vec{x}_i$ is an image

$\vec{y}_i = 1$ or $\vec{y}_i = 0$ if image i is a 5 or not, respectively

We fit a logistic-regression model

Digit classification



Newton's method

Motivation: Convex functions are often almost quadratic $f \approx f^2_{\vec{x}}$

Idea: Iteratively minimize quadratic approximation

$$f^2_{\vec{x}}(\vec{y}) := f(\vec{x}) + \nabla f(\vec{x})(\vec{y} - \vec{x}) + \frac{1}{2}(\vec{y} - \vec{x})^T \nabla^2 f(\vec{x})(\vec{y} - \vec{x}),$$

Minimum has closed form

$$\arg\min_{\vec{y} \in \mathbb{R}^n} f^2_{\vec{x}}(\vec{y}) = \vec{x} - \nabla^2 f(\vec{x})^{-1} \nabla f(\vec{x})$$

Proof

We have

$$\nabla f_{\vec{x}}^2 (y) = \nabla f (\vec{x}) + \nabla^2 f (\vec{x}) (\vec{y} - \vec{x})$$

It is equal to zero if

$$\nabla^2 f (\vec{x}) (\vec{y} - \vec{x}) = -\nabla f (\vec{x})$$

If the Hessian is positive definite, the only minimum of $f_{\vec{x}}^2$ is at

$$\vec{x} - \nabla^2 f (\vec{x})^{-1} \nabla f (\vec{x})$$

Newton's method

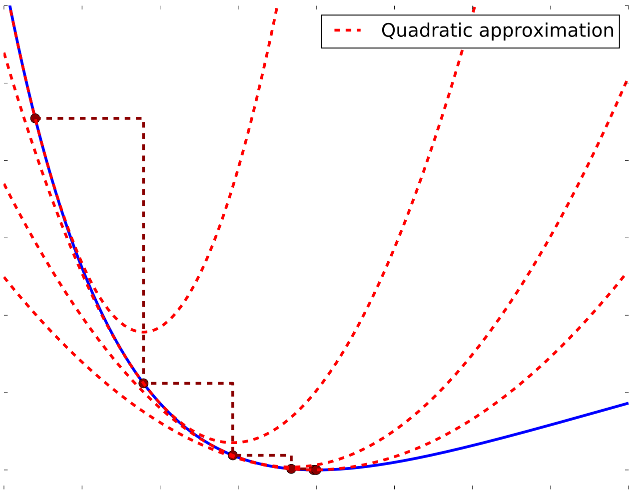
Set the initial point $\vec{x}^{(0)}$ to an arbitrary value

Update by setting

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \nabla^2 f\left(\vec{x}^{(k)}\right)^{-1} \nabla f\left(\vec{x}^{(k)}\right)$$

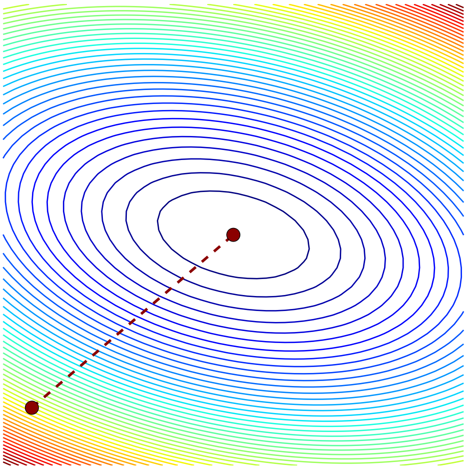
until a stopping criterion is met

Newton's method

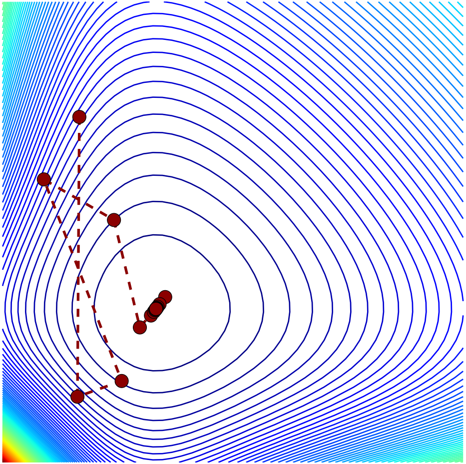


Quadratic function

Quadratic function



Convex function



Logistic regression

$$\frac{\partial^2 f(\vec{x})}{\partial \vec{x}[j] \partial \vec{x}[l]} = -\sum_{i=1}^{n} g(\langle \vec{x}^{(i)}, \vec{\beta} \rangle) \left(1 - g(\langle \vec{x}^{(i)}, \vec{\beta} \rangle)\right) \vec{x}^{(i)}[j] \, \vec{x}^{(i)}[l]$$

$$\nabla^2 f(\vec{\beta}) = -X^T G(\vec{\beta}) X$$

The rows of $X \in \mathbb{R}^{n \times p}$ contain $\vec{x}^{(1)}, \ldots \vec{x}^{(n)}$

G is a diagonal matrix such that

$$G(\vec{\beta})_{ii} := g(\langle \vec{x}^{(i)}, \vec{\beta} \rangle) \left(1 - g(\langle \vec{x}^{(i)}, \vec{\beta} \rangle)\right), \qquad 1 \leq i \leq n$$

Logistic regression

Newton updates are

$$\vec{\beta}^{(k+1)} := \vec{\beta}^{(k)} - \left(X^T G(\vec{\beta}^{(k)}) X \right)^{-1} \nabla f(\vec{\beta}^{(k)})$$

Sanity check: Cost function is concave, for any $\vec{\beta}, \vec{v} \in \mathbb{R}^p$

$$\vec{v}^T \nabla^2 f(\vec{\beta}) \vec{v} = -\sum_{i=1}^{n} G(\vec{\beta})_{ii} \left(X \vec{v} \right) [i]^2 \leq 0$$