Optimization in Machine Learning

Lecture 6: Calculus of Convexity, ML Examples, Sublevel sets and Epigraphs

Ganesh Ramakrishnan

Department of Computer Science
Dept of CSE, IIT Bombay
https://www.cse.iitb.ac.in/~ganesh

January, 2025

Outline of Content for Today on Convexity of Functions

1. Definition of Convexity, Strong Convexity and Strict Convexity [Done]
2. Examples of Convex Functions [Done]
3. Calculus of Convex Functions & More Properties [Almost Done]
4. Understanding the Convexity of Machine Learning Loss Functions
5. Direction Vector, Subgradients and Subdifferentials, Epigraphs and Sublevel sets,
6. First Order Convexity Conditions, Quasi Convexity
7. Basic Subgradient Calculus: Subgradients for non-differentiable convex functions
8. Convex Optimization Problems

[Recap] Composition with Vector Functions

- Composition of $g : \mathbb{R}^n \to \mathbb{R}^k$ and $h : \mathbb{R}^k \to \mathbb{R}$.

$$f(x) = h(g(x)) = h\left(g_1(x), \cdots, g_k(x)\right)$$

- f is convex if a) g_i's convex, h convex and non-decreasing in each argument or b) g_i concave, h convex and non-increasing in each argument
- Examples:
  - $f(x) = \sum_i \log(g_i(x))$ is concave if g_i is concave and positive
  - $\log \sum_{i=1}^k \exp(g_i(x))$ is convex if g_i is convex.

[Recap] Solution to Problem: Composition with Vector Functions

- Composition of $g : \mathbb{R}^n \to \mathbb{R}^k$ and $h : \mathbb{R}^k \to \mathbb{R}$.

$$f(x) = h(g(x)) = h(g_1(x), \cdots, g_k(x))$$

- f is convex if a) g_i's are convex, h convex and non-decreasing in each argument or b) g_i's are concave, h convex and non-increasing in each argument
- Examples:
  - $f(x) = \sum_i \log(g_i(x))$ is concave if g_i is concave and positive
  - $\log \sum_{i=1}^{k} \exp(g_i(x))$ is convex if g_i is convex. Hints?
    - ★ A function r is convex iff $r\left(\frac{x_1+x_2}{2}\right) \leq \frac{1}{2}r(x_1) + \frac{1}{2}r(x_2)$ (special case with $\alpha = \frac{1}{2}$ is equivalent to to the general case with $\alpha \in [0,1]$)

NOTE: FROM THE PROOF BELOW, IT APPEARS THAT LOG_SUM_EXP SHOULD BE STRICTLY CONVEX (equality holds in Cauchy Shwarz only when the x1 and x2 are the same). But STRONG CONVEXITY WHICH REQUIRES A QUADRATIC GAP LOOKS UNLIKELY

Using Midpoint convexity, let us prove that h(x) =log(sum(exp)) is convex (that it is non-increasing is easier)

$\log\left(\sum_i \exp(x^i)\right)$

To show

$$\underbrace{h\left(\frac{x_1}{2}+\frac{x_2}{2}\right)}_{\text{LHS}} \le \underbrace{\frac{1}{2}h(x_1)+\frac{1}{2}h(x_2)}_{\text{RHS}}$$

$$\text{LHS} = \log\left[\sum_i \exp\left(\frac{x_1^i}{2}+\frac{x_2^i}{2}\right)\right]$$

$$= \log\left[\sum_i \left(\exp(x_1^i)\right)^{1/2}\left(\exp(x_2^i)\right)^{1/2}\right]$$

$$= \log \sum_i \alpha_i \beta_i$$

$$\text{Let } \exp\left(\frac{x_1^i}{2}\right) = \alpha_i \qquad \exp\left(\frac{x_2^i}{2}\right) = \beta_i$$

$$\text{RHS} = \frac{1}{2}\log\left(\sum_i \exp(x_1^i)\right) + \frac{1}{2}\log\left(\sum_i \exp(x_2^i)\right)$$

$$= \log\left[\left(\sum_i \underbrace{\exp(x_1^i)}_{\alpha_i^2}\right)^{1/2}\left(\sum_i \underbrace{\exp(x_2^i)}_{\beta_i^2}\right)^{1/2}\right]$$

$$= \log\left[\left(\sum_i \alpha_i^2\right)^{1/2}\left(\sum_i \beta_i^2\right)^{1/2}\right]$$

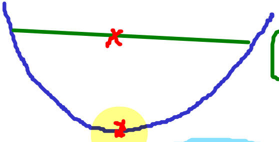
$$\boxed{\log \sum_i \alpha_i\beta_i \le \log\left[\left(\sum_i \alpha_i^2\right)^{1/2}\left(\sum_i \beta_i^2\right)^{1/2}\right]}$$

BY CAUCHY SHWARZ INEQUALITY & THE FACT THAT LOG IS MONOTONICALLY INCREASING FOR POSITIVE ARGUMENTS

$$\boxed{\sum_i \alpha_i\beta_i \le \left[\left(\sum_i \alpha_i^2\right)^{1/2}\left(\sum_i \beta_i^2\right)^{1/2}\right]}$$

A fn f is convex if $\forall\ \mu \in [0,1]$ & $\forall\ x_1, x_2 \in D$

$$f(\mu x_1 + (1-\mu)x_2) \leq \mu f(x_1) + (1-\mu) f(x_2)$$



Necessary and sufficient condition

MIDPOINT CONVEXITY
CONVEXITY AT 1/2

A fn f is convex if $\forall\ x_1, x_2 \in D$ (when $\mu = 1/2$) : $f\left(\frac{x_1}{2} + \frac{x_2}{2}\right) \leq \frac{f(x_1)}{2} + \frac{f(x_2)}{2}$

Tightens one for all

Proof sketch:

$$\mu x_1 + (1-\mu)x_2 \leq \tilde{x}_1/2 + \tilde{x}_2/2$$

H/W! Find $\tilde{x}_1, \tilde{x}_2$

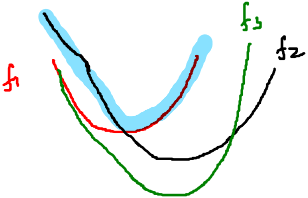
Pointwise Maximums and Supremums

Following functions are convex, but may not be differentiable everywhere.

- **Pointwise maximum:** If $f_1, f_2, \ldots, f_m$ are convex, then $f(\mathbf{x}) = max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$ is also convex. For example:
  - Sum of r largest components of $\mathbf{x} \in \Re^n$ $f(\mathbf{x}) = x_{[1]} + x_{[2]} + \ldots + x_{[r]}$, where $x_{[1]}$ is the i^{th} largest component of $\mathbf{x}$, is a convex function.
- **Pointwise supremum:** If $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ for every $\mathbf{y} \in \mathcal{S}$, then $g(\mathbf{x}) = \sup_{\mathbf{y} \in \mathcal{S}} f(\mathbf{x}, \mathbf{y})$ is convex. For example:
  - The function that returns the maximum eigenvalue of a symmetric matrix X, *viz.*, $\lambda_{max}(X) = \sup_{\mathbf{y} \in \mathcal{S}} \frac{\|X\mathbf{y}\|_2}{\|\mathbf{y}\|_2}$ is

Homework: Why is supremum required here?
Why wont max always suffice

- **Pointwise maximum:** If $f_1, f_2, \ldots, f_m$ are convex, then $f(\mathbf{x}) = max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$ is also convex. For example:
  - Sum of r largest components of $\mathbf{x} \in \Re^n$ $f(\mathbf{x}) = x_{[1]} + x_{[2]} + \ldots + x_{[r]}$, where $x_{[1]}$ is the i^{th} largest component of $\mathbf{x}$, is a convex function.



$(m = {}^nC_r)$

Linear functions are convex

$f_1(x) = v_1^T x$ $\quad$ $f_2(x) = v_2^T x$ $\quad$ $f_{{}^nC_r} = v_{{}^nC_r}^T x$

$v_1 = \underbrace{[0\ 1\ 0\ \ldots 1 \ldots 1\ \ 1\ 0]}_{\in \mathbb{R}^n}$

$v_2 = [0\ 0\ 1\ \ldots 1 \ldots 1\ \ 1\ 0] \in \mathbb{R}^n$

Pointwise Maximums and Supremums

$f(x) = \max_{v \in Perm([1\,1 \cdots 1\,0\,0\,0])} x^T v$ = sum of r largest components of x

$\underbrace{1\,1 \cdots 1}_{r\ 1\text{'s}}$

Linear & therefore convex

$\|x\|_\infty = \max\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

Following functions are convex, but may not be differentiable everywhere.

- **Pointwise maximum:** If $f_1, f_2, \ldots, f_m$ are convex, then $f(\mathbf{x}) = max\,\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$ is also convex. For example:
  - Sum of r largest components of $\mathbf{x} \in \Re^n$ $f(\mathbf{x}) = x_{[1]} + x_{[2]} + \ldots + x_{[r]}$, where $x_{[1]}$ is the i^{th} largest component of $\mathbf{x}$, is a convex function.
- **Pointwise supremum:** If $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ for every $\mathbf{y} \in \mathcal{S}$, then $g(\mathbf{x}) = \sup_{\mathbf{y} \in \mathcal{S}} f(\mathbf{x}, \mathbf{y})$ is convex. For example:
  - The function that returns the maximum eigenvalue of a symmetric matrix X, *viz.*, $\lambda_{max}(X) = \sup_{\mathbf{y} \in \mathcal{S}} \frac{\|X\mathbf{y}\|_2}{\|\mathbf{y}\|_2}$ is a convex function of the symmetrix matrix X.

Pointwise Maximums and Supremums

Following functions are convex, but may not be differentiable everywhere.

- **Pointwise maximum:** If $f_1, f_2, \ldots, f_m$ are convex, then $f(\mathbf{x}) = max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$ is also convex. For example:
  - Sum of r largest components of $\mathbf{x} \in \Re^n$ $f(\mathbf{x}) = x_{[1]} + x_{[2]} + \ldots + x_{[r]}$, where $x_{[1]}$ is the i^{th} largest component of $\mathbf{x}$, is a convex function.
- **Pointwise supremum:** If $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ for every $\mathbf{y} \in \mathcal{S}$, then $g(\mathbf{x}) = \sup\limits_{\mathbf{y} \in \mathcal{S}} f(\mathbf{x}, \mathbf{y})$ is convex. For example:
  - The function that returns the maximum eigenvalue of a symmetric matrix X, *viz.*, $\lambda_{max}(X) = \sup\limits_{\mathbf{y} \in \mathcal{S}} \frac{\|X\mathbf{y}\|_2}{\|\mathbf{y}\|_2}$ is a convex function of the symmetrix matrix X.

WHY: Symmetric matrix is diagonalizable.... using Eigenvector decompositon.

$\|X y\|_2^2 = y^T X^T X y$

Recall from PCA - the solution corresponds to the lambda in the direction of maximum variance obtained through the quadratic expansion

Some additional useful results concerning matrix norms

- Matrix Norm (for $A \in \Re^{m \times n}$) induced by vector norm N: $M_N(A) = \sup_{\mathbf{x} \neq 0} \frac{N(A\mathbf{x})}{N(\mathbf{x})}$

  Here, $\underset{s \in S}{sup}\ f(s) = \widehat{f}$ if $\widehat{f}$ is the minimum upper bound for $f(s)$ over $s \in S$.

  - *Eg.*: $M_N(I) = 1$ (*i.e.*, A is identity matrix I) irrespective of N
  - If $N = \|.\|_2$, $M_N(A) = \sqrt{\lambda_{max}(A^T A)} = \sigma_{max}(A)$ is the spectral norm, where $\lambda_{max}(A^T A)$ is the dominant eigenvalue of $A^T A$ and $\sigma_{max}(A)$ is called the largest singular value of A
  - $N = \|.\|_1 \implies M_N(A) = \max_j \sum_{i=1}^{m} |a_{ij}|$ & $N = \|.\|_\infty \implies M_N(A) = \max_i \sum_{j=1}^{n} |a_{ij}|$

With reference to previous slide, please note that for symmetric matrices, the singular values are the absolute values of the eigenvalues.

Some additional useful results concerning matrix norms

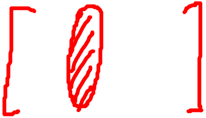
- Matrix Norm (for $A \in \Re^{m\times n}$) induced by vector norm N: $M_N(A) = \sup_{\mathbf{x}\neq 0} \frac{N(A\mathbf{x})}{N(\mathbf{x})}$

  Here, $\underset{s\in S}{sup}\ f(s) = \widehat{f}$ if $\widehat{f}$ is the minimum upper bound for $f(s)$ over $s \in S$.

  Homework: Read this and understand

  - *Eg.*: $M_N(I) = 1$ (*i.e.*, A is identity matrix I) irrespective of N
  - If $N = \|.\|_2$, $M_N(A) = \sqrt{\lambda_{max}(A^TA)} = \sigma_{max}(A)$ is the spectral norm, where $\lambda_{max}(A^TA)$ is the dominant eigenvalue of A^TA and $\sigma_{max}(A)$ is called the largest singular value of A
  - $N = \|.\|_1 \implies M_N(A) = \max_j \sum_{i=1}^{m} |a_{ij}|$ & $N = \|.\|_\infty \implies M_N(A) = \max_i \sum_{j=1}^{n} |a_{ij}|$

If ||x||_1 = 1 and we would like the 1 norm of the linear combination of the columns of A,
we would pick the x which has 1 only at the most important (or max) entry and 0 at other places
(since ||x||_1 is sum of absolutely values of x)
Hence makes sense for the peak to be where the 1 norm of A's columns is the highest



Column with maximum 1 norm

Some additional useful results concerning matrix norms

- Matrix Norm (for $A \in \Re^{m \times n}$) induced by vector norm N: $M_N(A) = \sup_{\mathbf{x} \neq 0} \frac{N(A\mathbf{x})}{N(\mathbf{x})}$

  Here, $\sup_{s \in S} f(s) = \widehat{f}$ if $\widehat{f}$ is the minimum upper bound for $f(s)$ over $s \in S$.
  - *Eg.*: $M_N(I) = 1$ (*i.e.*, A is identity matrix I) irrespective of N
  - If $N = \|.\|_2$, $M_N(A) = \sqrt{\lambda_{max}(A^T A)} = \sigma_{max}(A)$ is the spectral norm, where $\lambda_{max}(A^T A)$ is the dominant eigenvalue of $A^T A$ and $\sigma_{max}(A)$ is called the largest singular value of A
  - $N = \|.\|_1 \implies M_N(A) = \max_j \sum_{i=1}^{m} |a_{ij}|$ & $N = \|.\|_\infty \implies M_N(A) = \max_i \sum_{j=1}^{n} |a_{ij}|$
- The Schatten p-norm for $p \in [0,1]$ is another generalized norm defined by $\|A\|_{S_p} \equiv \|\sigma(A)\|_p$, where $\sigma(A)$ is the vector of singular values of A. Special cases are:
  - $p = 1 \implies$ Nuclear norm (or trace norm): $\|A\|_{S_1} = \sum_{i=1}^{min(m,n)} \sigma_i(A) = trace(\sqrt{A^T A}) = \|A\|_*$

    Recall (Lecture 2): Nuclear norm $\|A\|_*$ is the tightest convex lower bound of rank(A)
  - $p \to \infty \implies$ Spectral norm

Some additional useful results concerning matrix norms

- Matrix Norm (for $A \in \Re^{m\times n}$) induced by vector norm N: $M_N(A) = \sup\limits_{\mathbf{x}\neq 0} \frac{N(A\mathbf{x})}{N(\mathbf{x})}$

  Here, $\sup\limits_{s\in S} f(s) = \widehat{f}$ if $\widehat{f}$ is the minimum upper bound for $f(s)$ over $s \in S$.
  - *Eg.*: $M_N(I) = 1$ (*i.e.*, A is identity matrix I) irrespective of N
  - If $N = \|.\|_2$, $M_N(A) = \sqrt{\lambda_{max}(A^TA)} = \sigma_{max}(A)$ is the spectral norm, where $\lambda_{max}(A^TA)$ is the dominant eigenvalue of A^TA and $\sigma_{max}(A)$ is called the largest singular value of A
  - $N = \|.\|_1 \implies M_N(A) = \max\limits_j \sum\limits_{i=1}^{m} |a_{ij}|$ & $N = \|.\|_\infty \implies M_N(A) = \max\limits_i \sum\limits_{j=1}^{n} |a_{ij}|$
- The Schatten p-norm for $p \in [0,1]$ is another generalized norm defined by $\|A\|_{S_p} \equiv \|\sigma(A)\|_p$, where $\sigma(A)$ is the vector of singular values of A. Special cases are:
  - $p = 1 \implies$ Nuclear norm (or trace norm): $\|A\|_{S_1} = \sum\limits_{i=1}^{min(m,n)} \sigma_i(A) = trace(\sqrt{A^TA}) = \|A\|_*$

    Recall (Lecture 2): Nuclear norm $\|A\|_*$ is the tightest convex lower bound of rank(A)
  - $p \to \infty \implies$ Spectral norm

Which of the Following Loss Functions are Convex?

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i)) + \lambda \|\theta\|$

Which of the Following Loss Functions are Convex?

Recall that we proved convexity of logsumexp

$\sum \log(\exp(g_i(x)))$

$\exp(-y_i \text{ zero}^T x_i)$

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \underbrace{\log(1 + \exp(-y_i\theta^T x_i))}_{\text{LogSumExp}} + \lambda\|\theta\|$ YES

Proof: Generalization of the log(sum(exp)) proof where the g_i's are either 0 or linear functions

$g_i(x) = -y_i \theta^T x_i$

Long proof from first principles based on previous proof,
extended with more indices

Proof from first principles: $f(\theta) = \sum_{i=1}^{n} \underbrace{\overbrace{\log\left(\sum_{\theta} \exp(-y_i \theta^T x_i)\right)}^{g_i\,\theta}}_{g(\theta)=g_i(\theta)} + \underbrace{\lambda\|\theta\|}_{\lambda \geq 0 \text{ \& } \|.\|. \text{ is convex}}$

Need to prove convexity of $g_i(\theta)$

$$\underbrace{g_i(\mu\theta_1 + (1-\mu)\theta_2)}_{LHS} \overset{?}{\leq} \underbrace{\mu g_i(\theta_1) + (1-\mu) g_i(\theta_2)}_{RHS}$$

$$LHS = \log\left[\sum_{\theta} \exp\left(-y_i (\mu\theta_1 + (1-\mu)\theta_2)^T x_i\right)\right]$$

$$= \log\left[\sum_{\theta} \left(\exp(-y_i \theta_1^T x_i)\right)^{\mu_i} \left(\exp(-y_i \theta_2^T x_i)\right)^{(1-\mu_i)}\right]$$

$$(\text{if } \mu_i = \tfrac{1}{2}) = \log \sum_{\theta} \alpha_{i\theta} \beta_{i\theta}$$

Let $\exp\left(\frac{-y_i \theta_1^T x_i}{2}\right) = \alpha_{i\theta}$ $\exp\left(\frac{-y_i \theta_2^T x_i}{2}\right) = \beta_{i\theta}$

$$RHS = \mu \log\left[\sum_{\theta_1} \exp(-y_i \theta_1^T x_i)\right] + (1-\mu)\log\left(\sum_{\theta_2} \exp(-y_i \theta_2^T x_i)\right)$$

$$= \log\left[\left(\sum_{\theta_1} \exp(-y_i \theta_1^T x_i)\right)^{\mu_i} \left(\sum_{\theta_2} \exp(-y_i \theta_2^T x_i)\right)^{(1-\mu_i)}\right]$$

$$(\text{if } \mu_i = \tfrac{1}{2})$$

$$= \frac{1}{2} \log\left[\left(\sum_{\theta_1} \exp(-y_i \theta_1^T x_i)\right)\left(\sum_{\theta_2} \exp(-y_i \theta_2^T x_i)\right)\right]$$

$$= \log\left(\sum_{\theta} \alpha_{i\theta}^2\right)^{1/2} \left(\sum_{\theta} \beta_{i\theta}^2\right)^{1/2}$$

$$\log \sum_{\theta} \alpha_{i\theta}\beta_{i\theta} \le \log\left(\sum_{\theta} \alpha_{i\theta}^2\right)^{1/2}\left(\sum_{\theta} \beta_{i\theta}^2\right)^{1/2}$$

BECAUSE

$$\sum_{\theta} \alpha_{i\theta}\beta_{i\theta} \le \left(\sum_{\theta} \alpha_{i\theta}^2\right)^{1/2}\left(\sum_{\theta} \beta_{i\theta}^2\right)^{1/2}$$

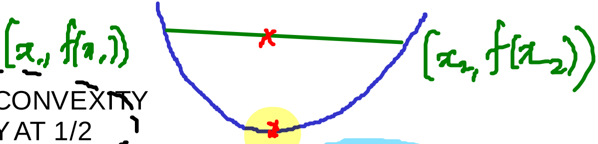
Implicitly, we have invoked that a function is convex iff it is midpoint convex

Let $\exp\left(\frac{-y_i\theta_1^T x_i}{2}\right) = \alpha_{i\theta}$ $\quad \exp\left(\frac{-y_i\theta_2^T x_i}{2}\right) = \beta_{i\theta}$

A fn f is convex if $\forall\ \mu \in [0,1]$ & $\forall\ x_1, x_2 \in D$

$$f(\mu x_1 + (1-\mu)x_2) \le \mu f(x_1) + (1-\mu) f(x_2)$$

Necessary and sufficient condition



MIDPOINT CONVEXITY
CONVEXITY AT 1/2

A fn f is convex if $\forall\ x_1, x_2 \in D$ (when $\mu = 1/2$):

$$f\left(\frac{x_1}{2} + \frac{x_2}{2}\right) \le \frac{f(x_1)}{2} + \frac{f(x_2)}{2}$$

Tightens one for all

Proof sketch: $\mu x_1 + (1-\mu)x_2 \le \tilde{x}_1/2 + \tilde{x}_2/2$

H/W! Find $\tilde{x}_1, \tilde{x}_2$

Which of the Following Loss Functions are Convex?

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i\theta^T x_i)) + \lambda\|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i\theta^T x_i\} + \lambda\|\theta\|$

Which of the Following Loss Functions are Convex?

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i\theta^T x_i)) + \lambda\|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \underbrace{\max\{0, 1 - y_i\theta^T x_i\}}_{\max\{\text{constant}, \text{linear}\}} + \lambda\|\theta\|$

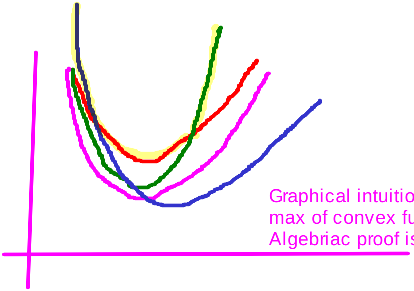
YES

Example of pointwise maximum of a finite number of functions

max{constant, linear}

convex

which is convex



Graphical intuition of why
max of convex functions are convex
Algebriac proof is much easier...

Which of the Following Loss Functions are Convex?

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i\theta^T x_i)) + \lambda\|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i\theta^T x_i\} + \lambda\|\theta\|$
- L1/L2 Reg Multi-class Logistic Regression:
  $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\} + \lambda \sum_{c=1}^{k} \|\theta_c\|$

Which of the Following Loss Functions are Convex?

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i)) + \lambda \|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i \theta^T x_i\} + \lambda \|\theta\|$
- L1/L2 Reg Multi-class Logistic Regression:
  $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\} + \lambda \sum_{c=1}^{k} \|\theta_c\|$

Generalization

YES:

LogSumExp with summations over lot more indicies

Which of the Following Loss Functions are Convex?

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i\theta^T x_i)) + \lambda\|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i\theta^T x_i\} + \lambda\|\theta\|$
- L1/L2 Reg Multi-class Logistic Regression: $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\} + \lambda \sum_{c=1}^{k} \|\theta_c\|$
- L1/L2 Reg Least Squares (Lasso): $L(\theta) = \sum_{i=1}^{n} (\theta^T x_i - y_i)^2 + \lambda\|\theta\|$

Which of the Following Loss Functions are Convex?

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i)) + \lambda \|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i \theta^T x_i\} + \lambda \|\theta\|$
- L1/L2 Reg Multi-class Logistic Regression: $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\} + \lambda \sum_{c=1}^{k} \|\theta_c\|$
- L1/L2 Reg Least Squares (Lasso): $L(\theta) = \sum_{i=1}^{n} \underline{(\theta^T x_i - y_i)}^2 + \lambda \|\theta\|$ YES

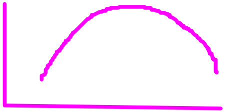
From first principles
OR
Based on composition of quadratic with an affine function

t_i^2 $\quad$ $t_i = \theta^T x_i - y_i$

Which of the Following Loss Functions are Convex?

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i)) + \lambda \|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i \theta^T x_i\} + \lambda \|\theta\|$
- L1/L2 Reg Multi-class Logistic Regression: $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\} + \lambda \sum_{c=1}^{k} \|\theta_c\|$
- L1/L2 Reg Least Squares (Lasso): $L(\theta) = \sum_{i=1}^{n} (\theta^T x_i - y_i)^2 + \lambda \|\theta\|$
- Matrix Completion: $L(X) = \sum_{i=1}^{n} ||y_i - A_i(X)||_2^2 + ||X||_*$

Which of the Following Loss Functions are Convex?



Is this concave?
Hint: PCA objective is also neither concave nor convex. Both have efficient solvers

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i\theta^T x_i)) + \lambda\|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i\theta^T x_i\} + \lambda\|\theta\|$
- L1/L2 Reg Multi-class Logistic Regression:
  $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\} + \lambda \sum_{c=1}^{k} \|\theta_c\|$
- L1/L2 Reg Least Squares (Lasso): $L(\theta) = \sum_{i=1}^{n} (\theta^T x_i - y_i)^2 + \lambda\|\theta\|$
- Matrix Completion: $L(X) = \sum_{i=1}^{n} \underbrace{||y_i - A_i(X)||_2^2}_{\text{Affine fn composed with } L_2 \text{ norm}} + ||X||_*$

YES

$||X||_* = \sum \sigma_i(X)$

Special case of Shatten Norm

$\|\sigma(X)\|_2$

$= \sup_{v \neq 0} \frac{\|Xv\|_2}{\|v\|_2}$

Example of pointwise supremum of infinite number of convex functions

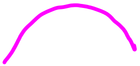
7 / 60

Which of the Following Loss Functions are Convex?

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i\theta^T x_i)) + \lambda\|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i\theta^T x_i\} + \lambda\|\theta\|$
- L1/L2 Reg Multi-class Logistic Regression: $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\} + \lambda \sum_{c=1}^{k} \|\theta_c\|$
- L1/L2 Reg Least Squares (Lasso): $L(\theta) = \sum_{i=1}^{n} (\theta^T x_i - y_i)^2 + \lambda\|\theta\|$
- Matrix Completion: $L(X) = \sum_{i=1}^{n} ||y_i - A_i(X)||_2^2 + ||X||_*$
- Soft-Max Contextual Bandits: $L(\theta) = \sum_{i=1}^{n} \frac{r_i}{p_i} \frac{\exp(\theta^T x_i^{a_i})}{\sum_{j=1}^{k} \exp(\theta^T x_i^j)} + \lambda\|\theta\|$

Which of the Following Loss Functions are Convex?

Recall: The Reg Logistic Regression loss is a cross entropy loss minimizing it is equivalent to maximizing The log likelihood objective



$\max_{\theta} \sum_{x_i} \log \left(\frac{\exp(y_i x_i^T \theta)}{1+\exp(y_i x_i^T \theta)} \right)$

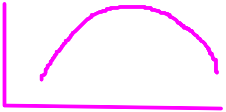
- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i)) + \lambda \|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i \theta^T x_i\} + \lambda \|\theta\|$
- L1/L2 Reg Multi-class Logistic Regression: $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\} + \lambda \sum_{c=1}^{k} \|\theta_c\|$
- L1/L2 Reg Least Squares (Lasso): $L(\theta) = \sum_{i=1}^{n} (\theta^T x_i - y_i)^2 + \lambda \|\theta\|$
- Matrix Completion: $L(X) = \sum_{i=1}^{n} ||y_i - A_i(X)||_2^2 + ||X||_*$

Convex but not with constraints

In fact, the PCA objective is also neither concave nor convex. But has efficient solvers

- Soft-Max Contextual Bandits: $L(\theta) = \sum_{i=1}^{n} \frac{r_i}{p_i} \frac{\exp(\theta^T x_i^{a_i})}{\sum_{j=1}^{k} \exp(\theta^T x_i^j)} + \lambda \|\theta\|$

Homework!



Inverse propensity estimate of the reward which we wanted to maximize
Here we should bother more about concavity than about convexity!

Relevance of Strong Convexity

- A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is strongly convex if there exists a $\mu > 0$ such that the function $g(x) = f(x) - \mu/2\|x\|^2$ is convex
- The parameter μ is the strong convexity parameter
- Geometrically, strong convexity means that there exists a quadratic lower bound on the growth of the function.
- Its easy to see that Strong Convexity implies Strict Convexity!
- Strong Convexity Doesn't imply the function is differentiable!
- If a function f is strongly convex and g is convex (not necessarily strongly convex), $f + g$ is strongly convex.
- $\|x\|^2$ is strongly convex!
- Hence for any convex function f, the function $f(x) + \lambda/2\|x\|^2$ is strongly convex!
- **Strong Convexity of a function results in faster convergence using certain descent algorithms.**

Relevance of Strong Convexity

- A function $f : \mathbb{R}^d \to \mathbb{R}$ is strongly convex if there exists a $\mu > 0$ such that the function $g(x) = f(x) - \mu/2\|x\|^2$ is convex
- The parameter μ is the strong convexity parameter
- Geometrically, strong convexity means that there exists a quadratic lower bound on the growth of the function.
- Its easy to see that Strong Convexity implies Strict Convexity!
- Strong Convexity Doesn't imply the function is differentiable!

Which means if a loss g(x)
is convex, we can derive
a strongly convex loss
(as we could on previous slide)
by adding an L2 norm
or strongly convex regularizer)

Eg: Hingeloss (convex but not differentiable) + L2 norm == Strongly convex

- If a function f is strongly convex and g is convex (not necessarily strongly convex), $f + g$ is strongly convex.
- $\|x\|^2$ is strongly convex!
- Hence for any convex function f, the function $f(x) + \lambda/2\|x\|^2$ is strongly convex!
- **Strong Convexity of a function results in faster convergence using certain descent algorithms.**

Outline of next few topics

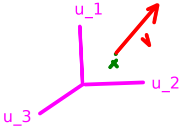
- Direction Vector, Directional derivative
- Quasi convexity & Sub-level sets of convex functions
- Convex Functions & their Epigraphs
- First-Order Convexity Conditions
- Subgradients, Subgradient Calculus and Convexity

The Direction Vector

- Consider a function $f(\mathbf{x})$, with $\mathbf{x} \in \Re^n$.
- We start with the concept of the direction at a point $\mathbf{x} \in \Re^n$.
- We will represent a vector by $\mathbf{x}$ and the k^{th} component of $\mathbf{x}$ by x_k.
- Let $\mathbf{u}^k$ be a unit vector pointing along the k^{th} coordinate axis in $\Re^n$;
- $u_k^k = 1$ and $u_j^k = 0, \ \forall j \neq k$
- An arbitrary direction vector $\mathbf{v}$ at $\mathbf{x}$ is a vector in $\Re^n$ with unit norm (*i.e.*, $||\mathbf{v}|| = 1$) and component v_k in the direction of $\mathbf{u}^k$.

The Direction Vector

- Consider a function $f(\mathbf{x})$, with $\mathbf{x} \in \Re^n$.
- We start with the concept of the direction at a point $\mathbf{x} \in \Re^n$.
- We will represent a vector by $\mathbf{x}$ and the k^{th} component of $\mathbf{x}$ by x_k.
- Let $\mathbf{u}^k$ be a unit vector pointing along the k^{th} coordinate axis in $\Re^n$;
- $u_k^k = 1$ and $u_j^k = 0$, $\forall j \neq k$ u's are called CARDINAL DIRECTIONS
- An arbitrary direction vector $\mathbf{v}$ at $\mathbf{x}$ is a vector in $\Re^n$ with unit norm ($i.e.$, $||\mathbf{v}|| = 1$) and component v_k in the direction of $\mathbf{u}^k$.



Directional derivative and the gradient vector

Let $f : \mathcal{D} \rightarrow \Re$, $\mathcal{D} \subseteq \Re^n$ be a function.

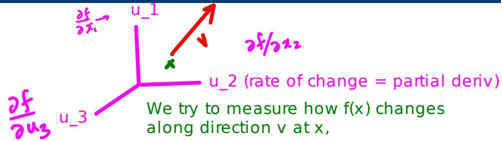
Definition

[Directional derivative]: The *directional derivative* of $f(\mathbf{x})$ at $\mathbf{x}$ in the direction of the unit vector $\mathbf{v}$ is

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{v}) - f(\mathbf{x})}{h} \tag{1}$$

provided the limit exists.

Directional derivative and the gradient vector



Let $f : \mathcal{D} \rightarrow \Re$, $\mathcal{D} \subseteq \Re^n$ be a function.

Definition

[Directional derivative]: The *directional derivative* of $f(\mathbf{x})$ at $\mathbf{x}$ in the direction of the unit vector $\mathbf{v}$ is

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{v}) - f(\mathbf{x})}{h} = v^T \underbrace{\nabla f(x)}_{\begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix}} \qquad (1)$$

provided the limit exists.

Directional Derivative

As a special case, when $\mathbf{v} = \mathbf{u}^k$ the directional derivative reduces to the partial derivative of f with respect to x_k.

$$D_{\mathbf{u}^k} f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_k}$$

Claim

If $f(\mathbf{x})$ is a differentiable function of $\mathbf{x} \in \Re^n$, then f has a directional derivative in the direction of any unit vector $\mathbf{v}$, and

$$D_{\mathbf{v}} f(\mathbf{x}) = \sum_{k=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_k} v_k = \nabla f^T v \tag{2}$$

Directional Derivative

As a special case, when $\mathbf{v} = \mathbf{u}^k$ the directional derivative reduces to the partial derivative of f with respect to x_k.

$$D_{\mathbf{u}^k} f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_k}$$

Claim

If $f(\mathbf{x})$ is a differentiable function of $\mathbf{x} \in \Re^n$, then f has a directional derivative in the direction of any unit vector $\mathbf{v}$, and

If f has min at x^*, then $D_v f(x^*)$ to be along an v? $D_v f(x^*) = 0 \;\forall v$

$\Updownarrow$

$\nabla f(x^*) = 0$

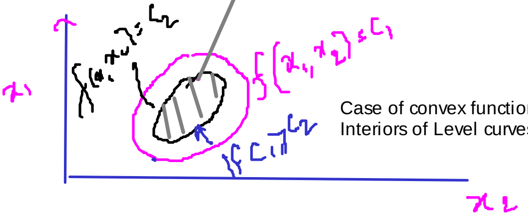
$$D_{\mathbf{v}} f(\mathbf{x}) = \sum_{k=1}^{n} \frac{\partial f(\mathbf{x})}{\partial x_k} v_k = \nabla f^T v \tag{2}$$

Assuming that as c reduced to below c1, the level curves only tend to keep being contained inside, the region inside is called a sub-level set

When moving within a pink level curve, f'(x) expected to be 0

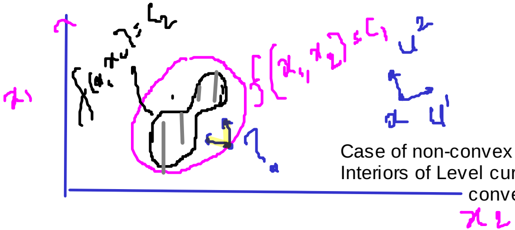
When moving within a black level curve, f'(x) expected to be 0

When moving from the pink level curve to the black level curve we do not expect f'(x) =0 (in fact, we expect f'(x) to be <= 0 (since c1 >= c2)



Case of convex functions
Interiors of Level curves expected to be convex

Motivations behind understanding and developing on directional derivatives
1) Design of algorithms and why certain algos work well for convex functions
2) Alternative definitions of convexity (in terms of first order and second order conditions)



Case of non-convex function
Interiors of Level curves might not be convex

In fact, the increasing or decreasing nature of f'(x) in a region is also closely connected to the convex/non-convex CURVATURE of the function

Sub-level Sets of Convex Functions

- Lets define *sub-level sets* of a convex function as follows:

Definition

[Sublevel Sets]: Let $\mathcal{D} \subseteq \Re^n$ be a nonempty set and $f : \mathcal{D} \rightarrow \Re$. The set

$$L_\alpha(f) = \{\mathbf{x} | \mathbf{x} \in \mathcal{D},\ f(\mathbf{x}) \leq \alpha\}$$

is called the $\alpha-$sub-level set of f.

Now if a function f is convex,

Sub-level Sets of Convex Functions

- Lets define *sub-level sets* of a convex function as follows:

Definition

[Sublevel Sets]: Let $\mathcal{D} \subseteq \Re^n$ be a nonempty set and $f : \mathcal{D} \to \Re$. The set

$$L_\alpha(f) = \{\mathbf{x} | \mathbf{x} \in \mathcal{D},\ f(\mathbf{x}) \leq \alpha\}$$

Interiors of Level curves expected to be convex

is called the $\alpha-$sub-level set of f.

Now if a function f is convex,

Assuming that as c reduced to below c1, the level curves only tend to keep being contained inside, the region inside is called a c1--sub-level set

Sub-level Sets of Convex Functions

- Lets define *sub-level sets* of a convex function as follows:

Definition

[Sublevel Sets]: Let $\mathcal{D} \subseteq \Re^n$ be a nonempty set and $f : \mathcal{D} \to \Re$. The set

$$L_\alpha(f) = \{\mathbf{x} | \mathbf{x} \in \mathcal{D},\ f(\mathbf{x}) \leq \alpha\}$$

is called the $\alpha-$sub-level set of f.

Now if a function f is convex, its $\alpha-$sub-level set is a convex set.

Sub-level Sets of Convex Functions

- Lets define *sub-level sets* of a convex function as follows:

Definition

[Sublevel Sets]: Let $\mathcal{D} \subseteq \Re^n$ be a nonempty set and $f : \mathcal{D} \to \Re$. The set

$$L_\alpha(f) = \{\mathbf{x} | \mathbf{x} \in \mathcal{D},\ f(\mathbf{x}) \leq \alpha\}$$

is called the $\alpha-$sub-level set of f.

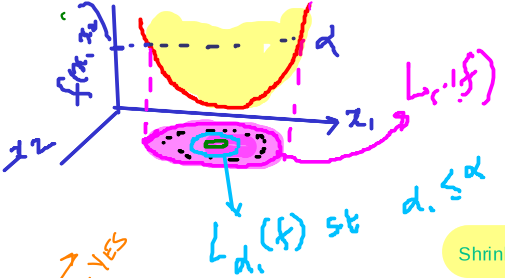
Now if a function f is convex, its $\alpha-$sub-level set is a convex set.

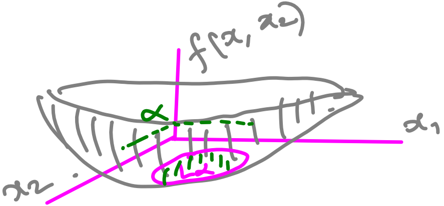
Proof

If $x_1, x_2 \in L_\alpha(f)$

$$f(\theta x_1 + (1-\theta)x_2) \leq \underbrace{\theta f(x_1)}_{\leq \theta\alpha} + \underbrace{(1-\theta) f(x_2)}_{\leq (1-\theta)\alpha} \leq \alpha$$

$\Rightarrow \theta x_1 + (1-\theta)x_2 \in L_\alpha(f)$





Shrinking α corresponds to downward movement along the function axis

Ans: YES

Q1: WHAT IF A THE EPIGRAPH IS CONVEX? DOES IT IMPLY THAT THE FUNCTION IS ALSO CONVEX?

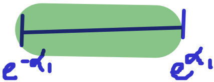
Q2: WHAT IF THE SUB-LEVEL SETS ARE CONVEX? DO THEY IMPLY THAT THE FUNCTION IS ALSO CONVEX?

Question: If all (or some) sub-level sets of a function are convex, is it implied that the function itself MUST be convex?

ANS: No.

Proof by counter-example
log |x| for $x>0$

$$\log|x| \le \alpha_1 \equiv |x| \le e^{\alpha_1}$$



Convex Function $\Rightarrow$ Convex Sub-level sets

Theorem

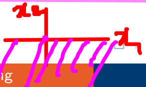
Let $\mathcal{D} \subseteq \Re^n$ be a nonempty convex set, and $f : \mathcal{D} \to \Re$ be a convex function. Then $L_\alpha(f)$ is a convex set for any $\alpha \in \Re$.

Proof: Consider $\mathbf{x}_1, \mathbf{x}_2 \in L_\alpha(f)$. Then by definition of the level set, $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$, $f(\mathbf{x}_1) \le \alpha$ and $f(\mathbf{x}_2) \le \alpha$. From convexity of $\mathcal{D}$ it follows that for all $\theta \in (0,1)$, $\mathbf{x} = \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathcal{D}$. Moreover, since f is also convex,

$$f(\mathbf{x}) \le \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2) \le \theta\alpha + (1-\theta)\alpha = \alpha$$

which implies that $\mathbf{x} \in L_\alpha(f)$. Thus, $L_\alpha(f)$ is a convex set. $\square$

The converse of this theorem does not hold. To illustrate this, consider the function $f(\mathbf{x}) = \frac{x_2}{1+2x_1^2}$. The 0-sublevel set of this function is $\{(x_1, x_2) \mid x_2 \le 0\}$, which is convex. However, the function $f(\mathbf{x})$ itself is not convex.



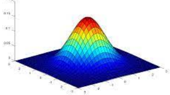
Convex Sub-level sets **DO NOT IMPLY** Convex Function

A function is called quasi-convex if all its sub-level sets are convex sets. Every quasi-convex function is not convex!

Consider the Negative of the normal distribution $-\frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. This function is quasi-convex but not convex.

: Show that the negative of the normal distribution is quasi-convex

Consider the simpler function $f(x) = -exp(-(x-\mu)^2)$.



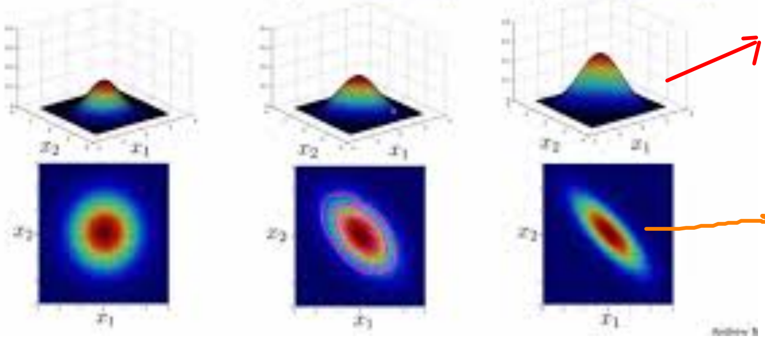
- Then $f'(x) = 2(x-\mu)exp(-(x-\mu)^2)$
- And
  $f''(x) = 2exp(-(x-\mu)^2) - 4(x-\mu)^2 exp(-(x-\mu)^2) = (2-4(x-\mu)^2)exp(-(x-\mu)^2)$
  which is < 0 if $(x-\mu)^2 > \frac{1}{2}$,
- Thus, the second derivative is negative if $x > \mu + \frac{1}{\sqrt{2}}$ or $x < -\mu - \frac{1}{\sqrt{2}}$.

  (Ignoring $\frac{1}{\sqrt{2\pi}}$, set $\sigma^2 = 1$)
- Recall from discussion of convexity of $f : \Re \to \Re$
  - The derivative is not non-decreasing everywhere $\implies$ function is not convex everywhere.

To prove that this function is quasi-convex, we can .... ????????????

Multivariate Gaussian (Normal) examples

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$



Not convex everywhere

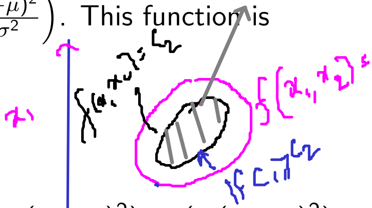
All sublevel sets are convex

Convex Sub-level sets DO NOT IMPLY Convex Function

A function is called quasi-convex if all its sub-level sets are convex sets. Every quasi-convex function is not convex!

Assuming that as c reduced to below c1, the level curves only tend to keep being contained inside, the region inside is called a c1–sub-level set

Consider the Negative of the normal distribution $-\frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. This function is quasi-convex but not convex.



Consider the simpler function $f(x) = -exp(-(x-\mu)^2)$.

- Then $f'(x) = 2(x-\mu)exp(-(x-\mu)^2)$
- And $f''(x) = 2exp(-(x-\mu)^2) - 4(x-\mu)^2 exp(-(x-\mu)^2) = (2-4(x-\mu)^2)exp(-(x-\mu)^2)$ which is < 0 if $(x-\mu)^2 > \frac{1}{2}$,
- Thus, the second derivative is negative if $x > \mu + \frac{1}{\sqrt{2}}$ or $x < -\mu - \frac{1}{\sqrt{2}}$.
- Recall from discussion of convexity of $f : \Re \rightarrow \Re$
  - The derivative is not non-decreasing everywhere $\Longrightarrow$ function is not convex everywhere.

To prove that this function is quasi-convex, we can ....