CS 453X: Class 10

Jacob Whitehill
Convex ML models
Convexity in higher dimensions

• For higher-dimensional \( f \), convexity is determined by the second derivative matrix, known as the **Hessian** of \( f \).

\[
H = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\
\frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2}
\end{bmatrix}
\]

• For \( f : \mathbb{R}^m \rightarrow \mathbb{R} \), \( f \) is convex if the Hessian matrix is positive semi-definite for every input \( \mathbf{x} \).
Positive semi-definite

- Positive semi-definite is the matrix analog of being “non-negative”.

- A real symmetric matrix \( A \) is **positive semi-definite (PSD)** if (equivalent conditions):

\[
x^T A x \geq 0
\]
Positive semi-definite

• Positive semi-definite is the matrix analog of being “non-negative”.

• A real symmetric matrix \( A \) is positive semi-definite (PSD) if (equivalent conditions):
  • All its eigenvalues are \( \geq 0 \).
  • If \( A \) happens to be diagonal, then its eigenvalues are the diagonal elements.
Positive semi-definite

- Positive semi-definite is the matrix analog of being “non-negative”.

- A real symmetric matrix $A$ is positive semi-definite (PSD) if (equivalent conditions):
  - All its eigenvalues are $\geq 0$.
    - If $A$ happens to be diagonal, then its eigenvalues are the diagonal elements.
  - For every vector $v$: $v^TAv \geq 0$
    - Therefore: If there exists any vector $v$ such that $v^TAv < 0$, then $A$ is not PSD.
Example

Suppose \( f(x, y) = 3x^2 + 2y^2 - 2 \).

Then the first derivatives are:

\[
\frac{\partial f}{\partial x} = 6x \quad \frac{\partial f}{\partial y} = 4y
\]

The Hessian matrix is therefore:

\[
H = \begin{bmatrix}
\frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\
\frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2}
\end{bmatrix} = \begin{bmatrix} 6 & 0 \\ 0 & 4 \end{bmatrix}
\]

Notice that \( H \) for this \( f \) does not depend on \((x,y)\).

Also, \( H \) is a diagonal matrix (with 6 and 4 on the diagonal). Hence, the eigenvalues are just 6 and 4. Since they are both non-negative, then \( f \) is convex.
Example

- Graph of $f(x, y) = 3x^2 + 2y^2 - 2$: 
Exercise

• Recall: if $H$ is the Hessian of $f$, then $f$ is convex if — at every $(x,y)$, we can show (equivalently):
  • $v^THv \geq 0$ for every $v$
  • All eigenvalues of $H$ are non-negative.

• Which of the following function(s) are convex?
  • $x^2 + y + 5$
  • $x^2 + 3xy$
  • $x^4 + xy + x^2$
Exercise

• Recall: if $H$ is the Hessian of $f$, then $f$ is convex if — at every $(x,y)$, we can show (equivalently):
  • $v^T Hv \geq 0$ for every $v$
  • All eigenvalues of $H$ are non-negative.

• Which of the following function(s) are convex?
  • $x^2 + y + 5$
  • $x^4 + xy + x^2$
Exercise

• Recall: if $H$ is the Hessian of $f$, then $f$ is convex if — at every $(x,y)$, we can show (equivalently):
  
  • $v^T Hv \geq 0$ for every $v$
  
  • All eigenvalues of $H$ are non-negative.

• Which of the following function(s) are convex?
  
  • $x^2 + y + 5$ \quad $H = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$ \quad Eigenvalues are 2, 0 => PSD.
  
  • $x^2 + 3xy$
  
  • $x^4 + xy + x^2$
Exercise

• Recall: if $H$ is the Hessian of $f$, then $f$ is convex if — at every $(x,y)$, we can show (equivalently):
  • $v^T H v \geq 0$ for every $v$
  • All eigenvalues of $H$ are non-negative.

• Which of the following function(s) are convex?
  • $x^2 + y + 5$
  • $x^2 + 3xy$
  • $H = \begin{bmatrix} 2 & 3 \\ 3 & 0 \end{bmatrix}$
  • $x^4 + xy + x^2$
Recall: if $H$ is the Hessian of $f$, then $f$ is convex if — at every $(x,y)$, we can show (equivalently):

- $v^T Hv \geq 0$ for every $v$
- All eigenvalues of $H$ are non-negative.

Which of the following function(s) are convex?

- $x^2 + y + 5$
- $x^2 + 3xy$
- $x^4 + xy + x^2$

$$H = \begin{bmatrix} 2 & 3 \\ 3 & 0 \end{bmatrix} \quad v = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad v^T Hv = -4$$

Not PSD.
Exercise

• Recall: if $H$ is the Hessian of $f$, then $f$ is convex if — at every $(x,y)$, we can show (equivalently):
  
  • $v^T Hv \geq 0$ for every $v$
  • All eigenvalues of $H$ are non-negative.

• Which of the following function(s) are convex?
  
  • $x^2 + y + 5$
  • $x^2 + 3xy$
  • $x^4 + xy + x^2$
  
  $H = \begin{bmatrix} 12x^2 + 2 & 1 \\ 1 & 0 \end{bmatrix}$
Exercise

• Recall: if $H$ is the Hessian of $f$, then $f$ is convex if — at every $(x,y)$, we can show (equivalently):
  • $v^T Hv \geq 0$ for every $v$
  • All eigenvalues of $H$ are non-negative.

• Which of the following function(s) are convex?
  • $x^2 + y + 5$
  • $x^2 + 3xy$
  • $x^4 + xy + x^2$

\[
\begin{bmatrix}
12x^2 + 2 & 1 \\
1 & 0
\end{bmatrix}
\begin{bmatrix}
\frac{-1}{15}
\end{bmatrix}
\begin{bmatrix}
\frac{-1}{15}
\end{bmatrix} = -16
\]

Not PSD.
Convexity of linear regression and softmax regression

- Why are they convex?

- First, recall that, for any matrices $A, B$ that can be multiplied:

  - $(AB)^T = B^TA^T$
Convexity of linear regression and softmax regression

- Why are they convex?

- Next, recall the gradient of $f_{\text{MSE}}$ (for linear regression):

$$\nabla_w f_{\text{MSE}} = X(\hat{y} - y)$$

$$= X(X^T w - y)$$

$$H = XX^T$$
Convexity of linear regression and softmax regression

• Why are they convex?

• Next, recall the gradient of $f_{\text{MSE}}$ (for linear regression):

\[
\nabla_w f_{\text{MSE}} = X(\hat{y} - y) \\
= X(X^T w - y) \\
H = XX^T
\]

• For any vector $v$, we have:

\[
v^T XX^T v = (X^T v)^T (X^T v) \\
\geq 0
\]
Convex ML models

• Beyond linear regression and softmax regression, what other convex ML models are there?

• One of the most prominent is the support vector machine (SVM).
Constrained optimization
Unconstrained optimization

- So far, the ML methods we have examined are based on optimizing some **objective function** (loss or accuracy).

- The optimization variable has been **unconstrained** — it can be any value in $\mathbb{R}^m$.

- Unconstrained optimal solutions exist at critical points of the objective function $f$, i.e., where the gradient of $f$ is 0, e.g.:

  - The minimum of this function is at $x=0$. 
Constrained optimization

- Things become more complicated when we put a constraint on the optimization variables.

- What if we want to minimize $f$ subject to the inequality constraint that $x \geq 1$?
Constrained optimization

• Things become more complicated when we put a constraint on the optimization variables.

• What if we want to minimize $f$ subject to the inequality constraint that $x \geq 1$?

• The solution no longer occurs at a critical point of $f$.

• The minimum of $f$, constrained s.t. $x \geq 1$, is at $x=1$. 

[Diagram showing the inequality constraint and the minimum point at x=1]
Constrained optimization methods

- A variety of techniques exist for solving constrained optimization problems.

- Many of these are applicable when the objective function $f$ is convex.

- Two widely used techniques:
  - Lagrange multipliers
  - Karush-Kuhn-Tucker (KKT) optimality conditions
Lagrange multipliers
Lagrange multipliers

- Lagrange multipliers are useful for solving optimization problems involving *equality constraints*, e.g., minimize:

\[ f(x, y) = x^2 + 3y^2 \text{ subject to } x + y = 2 \]
Lagrange multipliers

- Lagrange multipliers are useful for solving optimization problems involving equality constraints, e.g., minimize:

\[ f(x, y) = x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \]

Objective function
Equality constraint

![3D plot of the function and contour plot with the constraint graphed]
Lagrange multipliers

• We can express the equality constraint \((x+y=2)\) as a constraint function \(g\).

• We define \(g\) so that \(g(x,y) = 0\) when the constraint is satisfied:

\[
g(x, y) = \text{?}
\]
Lagrange multipliers

- We can express the equality constraint \((x+y=2)\) as a constraint function \(g\).

- We define \(g\) so that \(g(x,y) = 0\) when the constraint is satisfied:

\[
g(x, y) = x + y - 2
\]
Lagrange multipliers

- To solve the constrained optimization problem, we define the Lagrangian function $L$ in terms of:
  - The original optimization variables.
  - The Lagrange multiplier(s) $\alpha$ (one for each constraint).
- For one constraint $g$, we have:
  \[
  L(x, y, \alpha) = f(x, y) + \alpha g(x, y)
  \]
Lagrange multipliers

- The solution occurs at a critical point of $L$, i.e., where the derivative of $L$ with respect to $x$, $y$, and $\alpha = 0$.

\[
L(x, y, \alpha) = f(x, y) + \alpha g(x, y)
\]

\[
\frac{\partial L}{\partial x} = 0
\]

\[
\frac{\partial L}{\partial y} = 0
\]

\[
\frac{\partial L}{\partial \alpha} = 0
\]
Example

\[ f(x, y) = x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \]
Example

\[ f(x, y) = x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \]
\[ L(x, y, \alpha) = x^2 + 3y^2 + \alpha(x + y - 2) \]
Example

\[
\begin{align*}
  f(x, y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\
  L(x, y, \alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \\
  \frac{\partial L}{\partial x} &= 2x + \alpha = 0 \\
  \frac{\partial L}{\partial y} &= 6y + \alpha = 0 \\
  \frac{\partial L}{\partial \alpha} &= x + y - 2 = 0
\end{align*}
\]
Example

\[ f(x, y) = x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \]

\[ L(x, y, \alpha) = x^2 + 3y^2 + \alpha(x + y - 2) \]

\[
\frac{\partial L}{\partial x} = 2x + \alpha = 0
\]
\[
\frac{\partial L}{\partial y} = 6y + \alpha = 0
\]
\[
\frac{\partial L}{\partial \alpha} = x + y - 2 = 0
\]
\[
2x = 6y
\]
Example

\[ f(x, y) = x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \]

\[ L(x, y, \alpha) = x^2 + 3y^2 + \alpha(x + y - 2) \]

\[ \frac{\partial L}{\partial x} = 2x + \alpha = 0 \]

\[ \frac{\partial L}{\partial y} = 6y + \alpha = 0 \]

\[ \frac{\partial L}{\partial \alpha} = x + y - 2 = 0 \]

\[ 2x = 6y \]

\[ x = 3y \]
Example

\[ f(x, y) = x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \]

\[ L(x, y, \alpha) = x^2 + 3y^2 + \alpha(x + y - 2) \]

\[
\frac{\partial L}{\partial x} = 2x + \alpha = 0 \\
\frac{\partial L}{\partial y} = 6y + \alpha = 0 \\
\frac{\partial L}{\partial \alpha} = x + y - 2 = 0 \\
2x = 6y \\
x = 3y \\
3y + y - 2 = 0 \]
Example

\[ f(x, y) = x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \]

\[ L(x, y, \alpha) = x^2 + 3y^2 + \alpha(x + y - 2) \]

\[ \frac{\partial L}{\partial x} = 2x + \alpha = 0 \]

\[ \frac{\partial L}{\partial y} = 6y + \alpha = 0 \]

\[ \frac{\partial L}{\partial \alpha} = x + y - 2 = 0 \]

\[ 2x = 6y \]

\[ x = 3y \]

\[ 3y + y - 2 = 0 \]

\[ 4y = 2 \]
Example

\[ f(x, y) = x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \]

\[ L(x, y, \alpha) = x^2 + 3y^2 + \alpha(x + y - 2) \]

\[ \frac{\partial L}{\partial x} = 2x + \alpha = 0 \]

\[ \frac{\partial L}{\partial y} = 6y + \alpha = 0 \]

\[ \frac{\partial L}{\partial \alpha} = x + y - 2 = 0 \]

\[ 2x = 6y \]

\[ x = 3y \]

\[ 3y + y - 2 = 0 \]

\[ 4y = 2 \]

\[ y = 1/2 \]
Example

\[
f(x, y) = x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2
\]

\[
L(x, y, \alpha) = x^2 + 3y^2 + \alpha(x + y - 2)
\]

\[
\frac{\partial L}{\partial x} = 2x + \alpha = 0
\]

\[
\frac{\partial L}{\partial y} = 6y + \alpha = 0
\]

\[
\frac{\partial L}{\partial \alpha} = x + y - 2 = 0
\]

\[
2x = 6y
\]

\[
x = 3y
\]

\[
3y + y - 2 = 0
\]

\[
4y = 2
\]

\[
y = 1/2
\]

\[
x = 3/2
\]
Exercise

• Minimize:

\[ f(x, y) = x + y \quad \text{subject to} \quad x^2 + y^2 = 1 \]
Exercise

- Minimize:

\[ f(x, y) = x + y \] subject to \[ x^2 + y^2 = 1 \]
Exercise

• Minimize:

\[ f(x, y) = x + y \quad \text{subject to} \quad x^2 + y^2 = 1 \]

\[ L(x, y, \alpha) = x + y + \alpha(x^2 + y^2 - 1) \]
Exercise

• Minimize:

\[
\begin{align*}
f(x, y) &= x + y \quad \text{subject to} \quad x^2 + y^2 = 1 \\
L(x, y, \alpha) &= x + y + \alpha(x^2 + y^2 - 1) \\
\frac{\partial L}{\partial x} &= 1 + 2\alpha x = 0 \\
\frac{\partial L}{\partial y} &= 1 + 2\alpha y = 0 \\
\frac{\partial L}{\partial \alpha} &= x^2 + y^2 - 1 = 0
\end{align*}
\]
Exercise

• Minimize:

\[ f(x, y) = x + y \text{ subject to } x^2 + y^2 = 1 \]

\[ L(x, y, \alpha) = x + y + \alpha(x^2 + y^2 - 1) \]

\[ \frac{\partial L}{\partial x} = 1 + 2\alpha x = 0 \]

\[ \frac{\partial L}{\partial y} = 1 + 2\alpha y = 0 \]

\[ \frac{\partial L}{\partial \alpha} = x^2 + y^2 - 1 = 0 \]

\[ 2\alpha x = -1 \]

\[ x = -1/(2\alpha) \]

\[ y = -1/(2\alpha) = x \]

\[ x^2 + (x)^2 - 1 = 0 \]

\[ 2x^2 = 1 \]

\[ x^2 = 1/2 \]

\[ x = y = \pm 1/\sqrt{2} \]
Exercise

• Try \( x = y = +1/\sqrt{2} \): \( f(+1/\sqrt{2}, +1/\sqrt{2}) = +2/\sqrt{2} = +\sqrt{2}/2 \)  Maximum

• Try \( x = y = -1/\sqrt{2} \): \( f(-1/\sqrt{2}, -1/\sqrt{2}) = -2/\sqrt{2} = -\sqrt{2}/2 \) Minimum
Support vector machines
Support vector machines (SVMs) are a ML model for binary classification.

SVMs are optimized using constrained optimization rather than unconstrained optimization (e.g., for logistic regression).
Support vector machines

- Suppose we have the following set of training data (blue is negative, red is positive):

- Examples above the line will be classified as positive; examples below the line will be classified as negative.

- Which line (or hyperplane in higher dimensions) would likely perform better on testing data, and why?
Support vector machines

- For any hyperplane $H$ that perfectly separates the positive from the negative examples:
  - Find the subset $S^-$ of negative examples that lie closest to $H$.
  - The points in $S^-$ lie in a hyperplane $H^-$ parallel to $H$.
  - Denote the shortest distance between $H^-$ and $H$ as $d^-$. 
Support vector machines

- For any hyperplane $H$ that perfectly separates the positive from the negative examples:
  - Find the subset $S^+$ of + examples that lie closest to $H$.
  - The points in $S^+$ lie in a hyperplane $H^+$ parallel to $H$.
  - Denote the shortest distance between $H^+$ and $H$ as $d^+$. 
For any hyperplane $H$ that perfectly separates the positive from the negative examples:

- Let $d$ denote the **margin** — the sum of $d^+$ and $d^-$.

- The optimization objective of SVMs is to find a separating hyperplane $H$ that maximizes $d$. 

Support vector machines