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*.

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\[2.2ex] \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\[2.2ex] \vdots & \vdots & \ddots & \vdots \\[2.2ex] \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

- For $f : \mathbb{R}^m \to \mathbb{R}$, *f* is convex if the Hessian matrix is positive semi-definite for *every* input **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):

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 ≥ 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 ≥ 0.
    - If **A** happens to be diagonal, then its eigenvalues are the diagonal elements.
  - For every vector **v**: $\mathbf{v}^{\mathsf{T}}\mathbf{A}\mathbf{v} \geq 0$
    - Therefore: If there exists *any* vector **v** such that $\mathbf{v}^{\mathsf{T}}\mathbf{A}\mathbf{v} < 0$, then **A** is *not* PSD.

Example

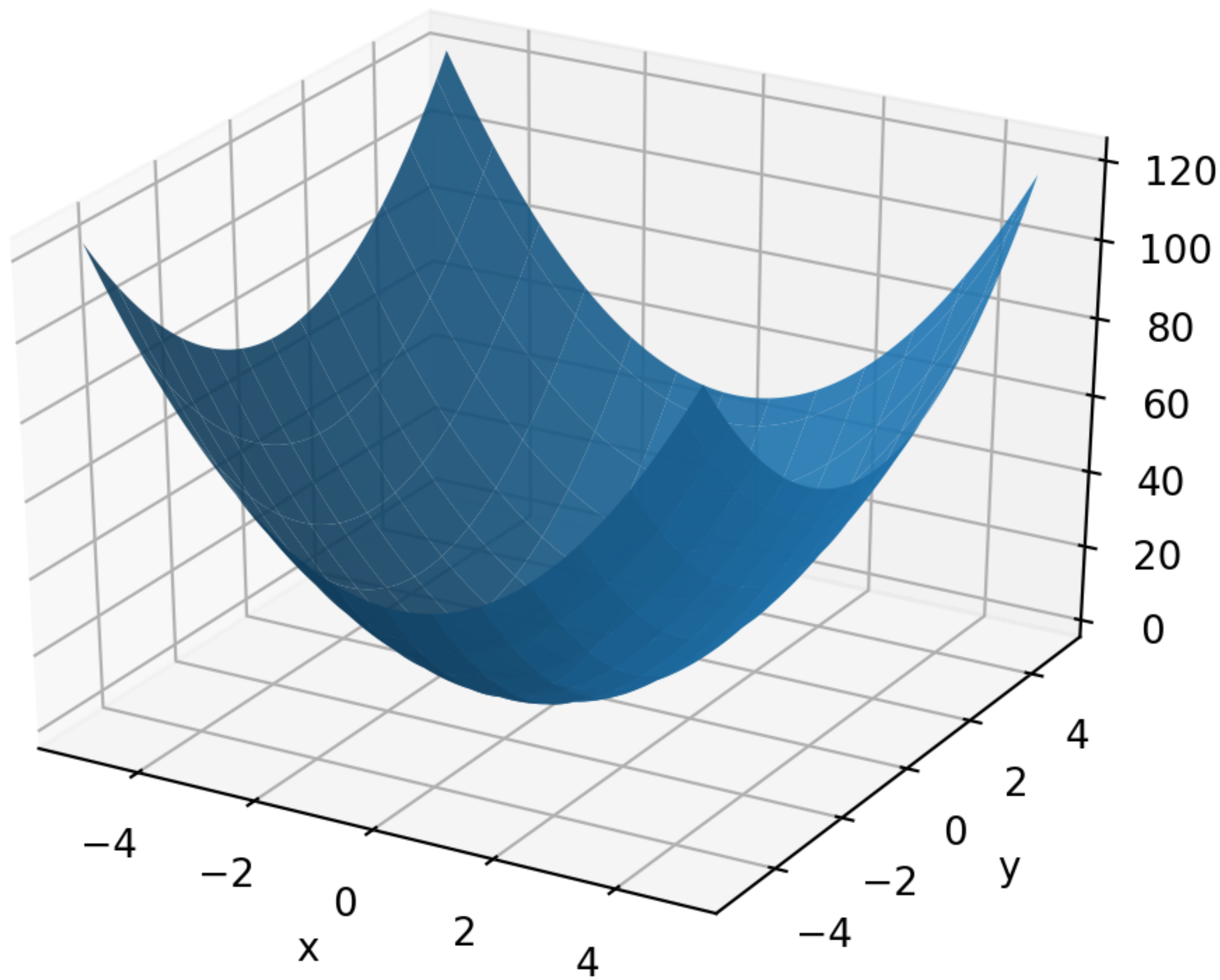
- Suppose $f(x, y) = 3x^2 + 2y^2 - 2$.
- Then the first derivatives are: $\frac{\partial f}{\partial x} = 6x \quad \frac{\partial f}{\partial y} = 4y$
- The Hessian matrix is therefore:

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

- 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):
    - $\mathbf{v}^T\mathbf{H}\mathbf{v} \geq 0$ for *every* **v**
    - All eigenvalues of **H** are non-negative.

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

Exercise

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

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

Exercise

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

- Which of the following function(s) are convex?
  - $x^2 + y + 5$ $\quad \mathbf{H} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$ **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):
  - $\mathbf{v}^T\mathbf{H}\mathbf{v} \geq 0$ for *every* **v**
  - All eigenvalues of **H** are non-negative.

- Which of the following function(s) are convex?

  - $x^2 + y + 5$
  - $x^2 + 3xy$ $\quad \mathbf{H} = \begin{bmatrix} 2 & 3 \\ 3 & 0 \end{bmatrix}$
  - $x^4 + xy + x^2$

Exercise

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

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

    Not PSD.
  - $x^4 + xy + x^2$

Exercise

- Recall: if **H** is the Hessian of *f*, then *f* is convex if — at *every* (x,y), we can show (equivalently):
  - $\mathbf{v}^\mathsf{T}\mathbf{H}\mathbf{v} \geq 0$ for *every* **v**
  - All eigenvalues of **H** are non-negative.
- Which of the following function(s) are convex?
  - $x^2 + y + 5$
  - $x^2 + 3xy$
  - $x^4 + xy + x^2$ $\quad \mathbf{H} = \begin{bmatrix} 12x^2 + 2 & 1 \\ 1 & 0 \end{bmatrix}$

Exercise

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

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

$x = 1$

$$\mathbf{H} = \begin{bmatrix} 12x^2 + 2 & 1 \\ 1 & 0 \end{bmatrix} \quad \mathbf{v} = \begin{bmatrix} -1 \\ 15 \end{bmatrix} \quad \mathbf{v}^\top\mathbf{H}\mathbf{v} = -16$$

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:
    - $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$

Convexity of linear regression and softmax regression

- Why are they convex?

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

$$
\begin{aligned}
\nabla_{\mathbf{w}} f_{\text{MSE}} &= \mathbf{X}(\hat{\mathbf{y}} - \mathbf{y}) \\
&= \mathbf{X}(\mathbf{X}^\top \mathbf{w} - \mathbf{y}) \\
\mathbf{H} &= \mathbf{X}\mathbf{X}^\top
\end{aligned}
$$

Convexity of linear regression and softmax regression

- Why are they convex?

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

$$
\begin{aligned}
\nabla_{\mathbf{w}} f_{\text{MSE}} &= \mathbf{X}(\hat{\mathbf{y}} - \mathbf{y}) \\
&= \mathbf{X}(\mathbf{X}^\top \mathbf{w} - \mathbf{y}) \\
\mathbf{H} &= \mathbf{X}\mathbf{X}^\top
\end{aligned}
$$

- For any vector **v**, we have:

$$
\begin{aligned}
\mathbf{v}^\top \mathbf{X}\mathbf{X}^\top \mathbf{v} &= \left(\mathbf{X}^\top \mathbf{v}\right)^\top \left(\mathbf{X}^\top \mathbf{v}\right) \\
&\geq 0
\end{aligned}
$$

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)**.

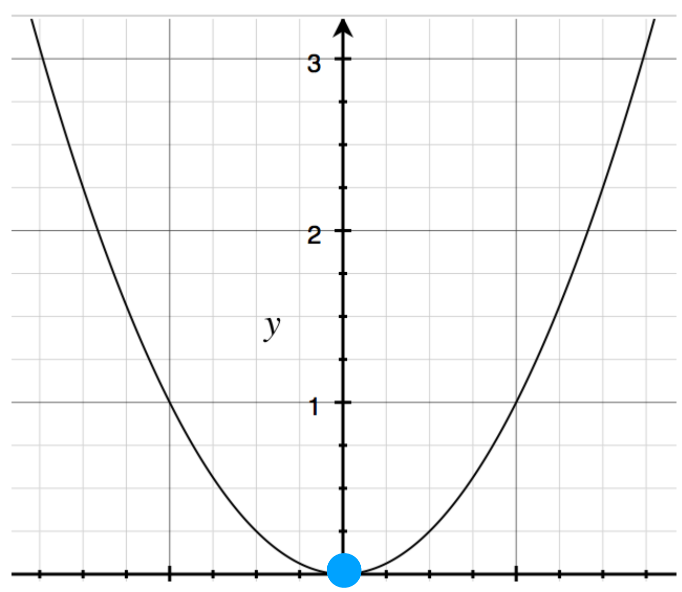
Constrained optimization

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.:

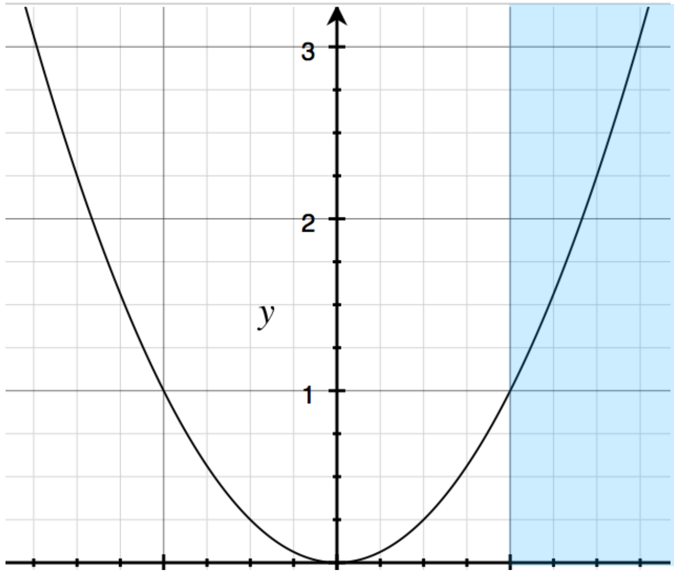


  - The minimum of this function is at x=0.

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 $x \geq 1$?



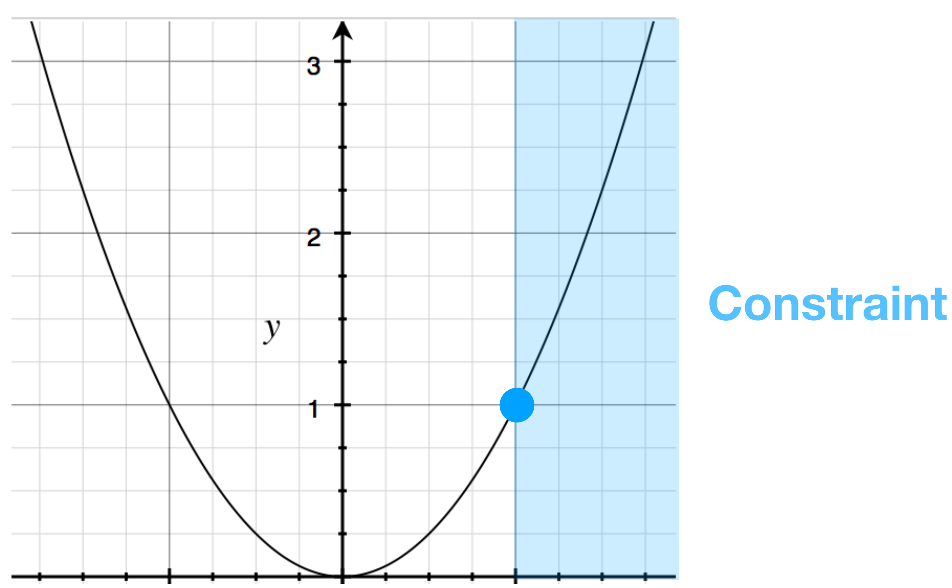
Constraint

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.

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

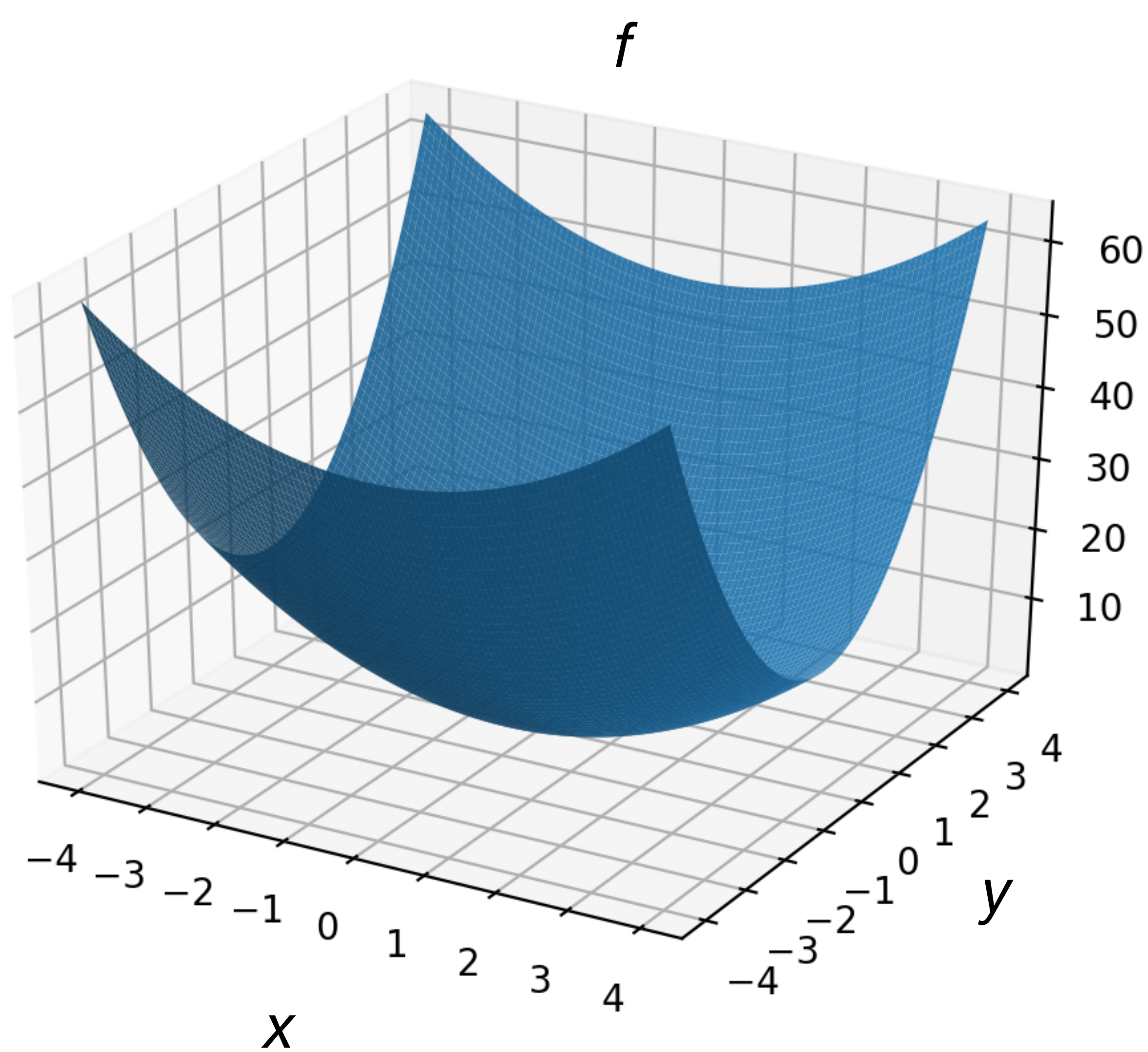
  - Karush-Kuhn-Tucker (KKT) optimality conditions

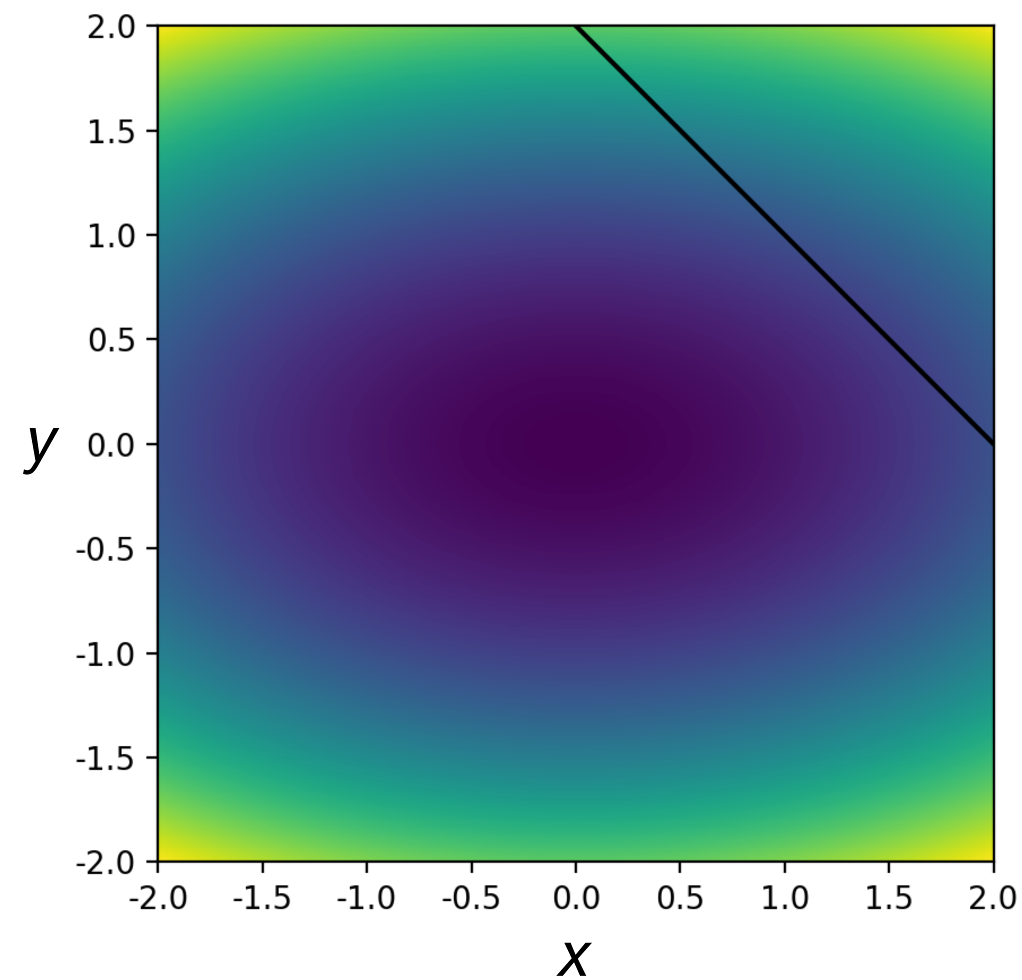
Lagrange multipliers

Lagrange multipliers

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

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





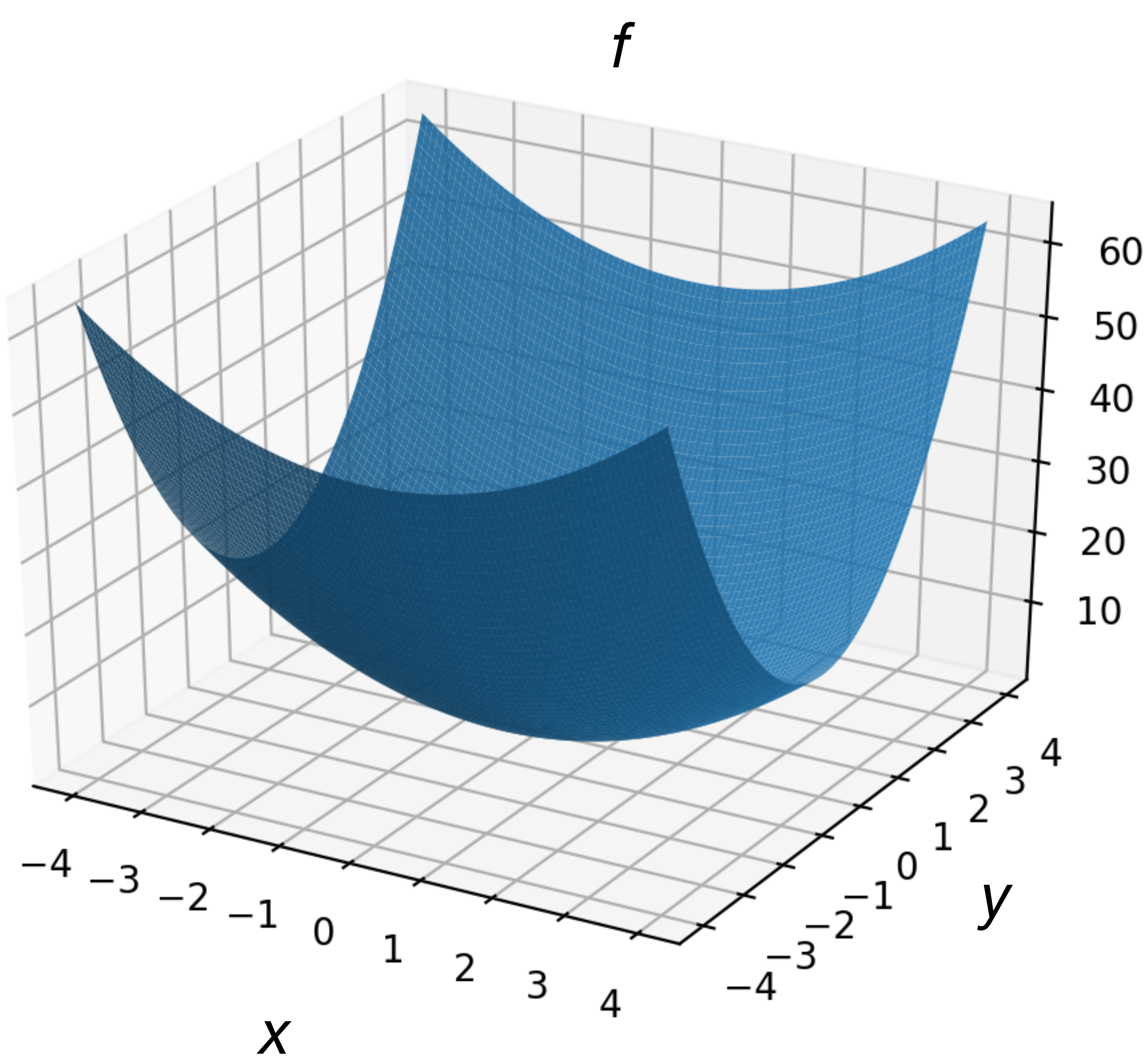
Lagrange multipliers

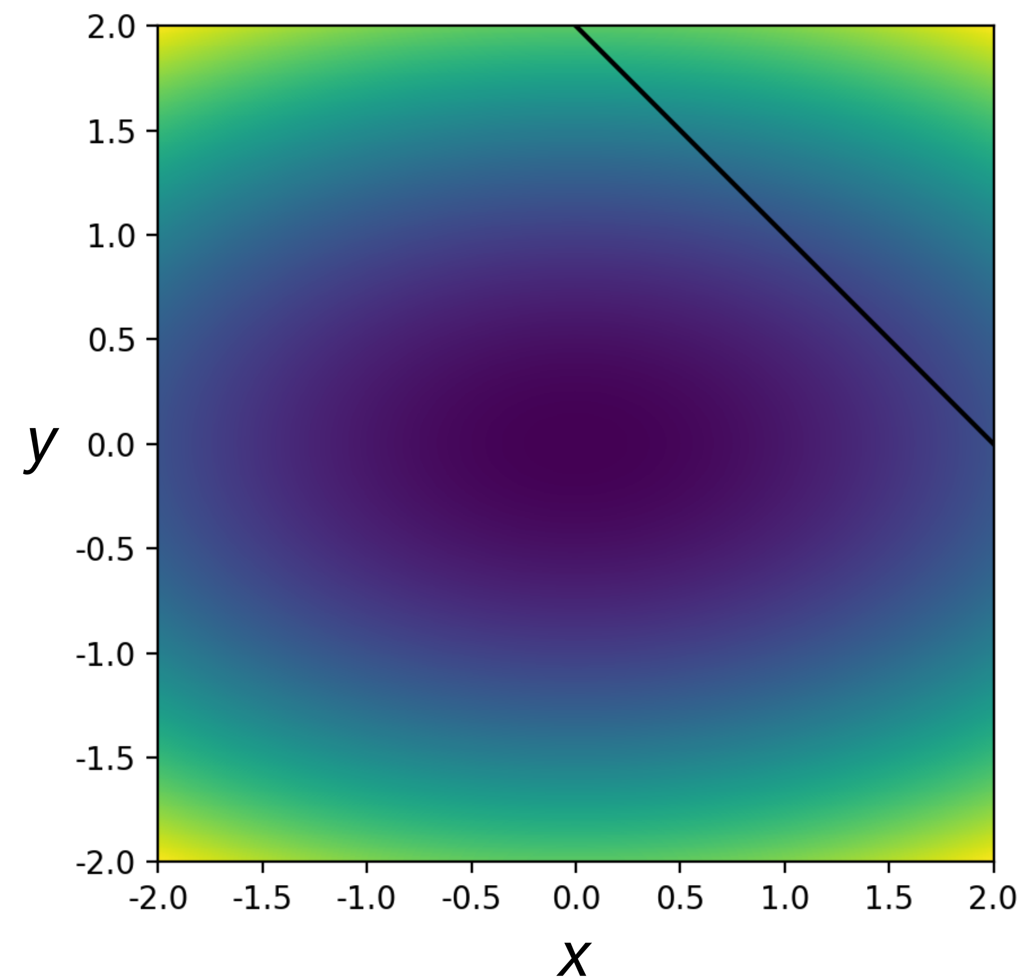
- Lagrange multipliers are useful for solving optimization problems involving **equality constraints**, e.g., minimize:

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

Objective function

Equality constraint





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:

$$g(x, y) = \boxed{?}$$

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) α (one for each constraint).
- For one constraint *g*, we have:

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

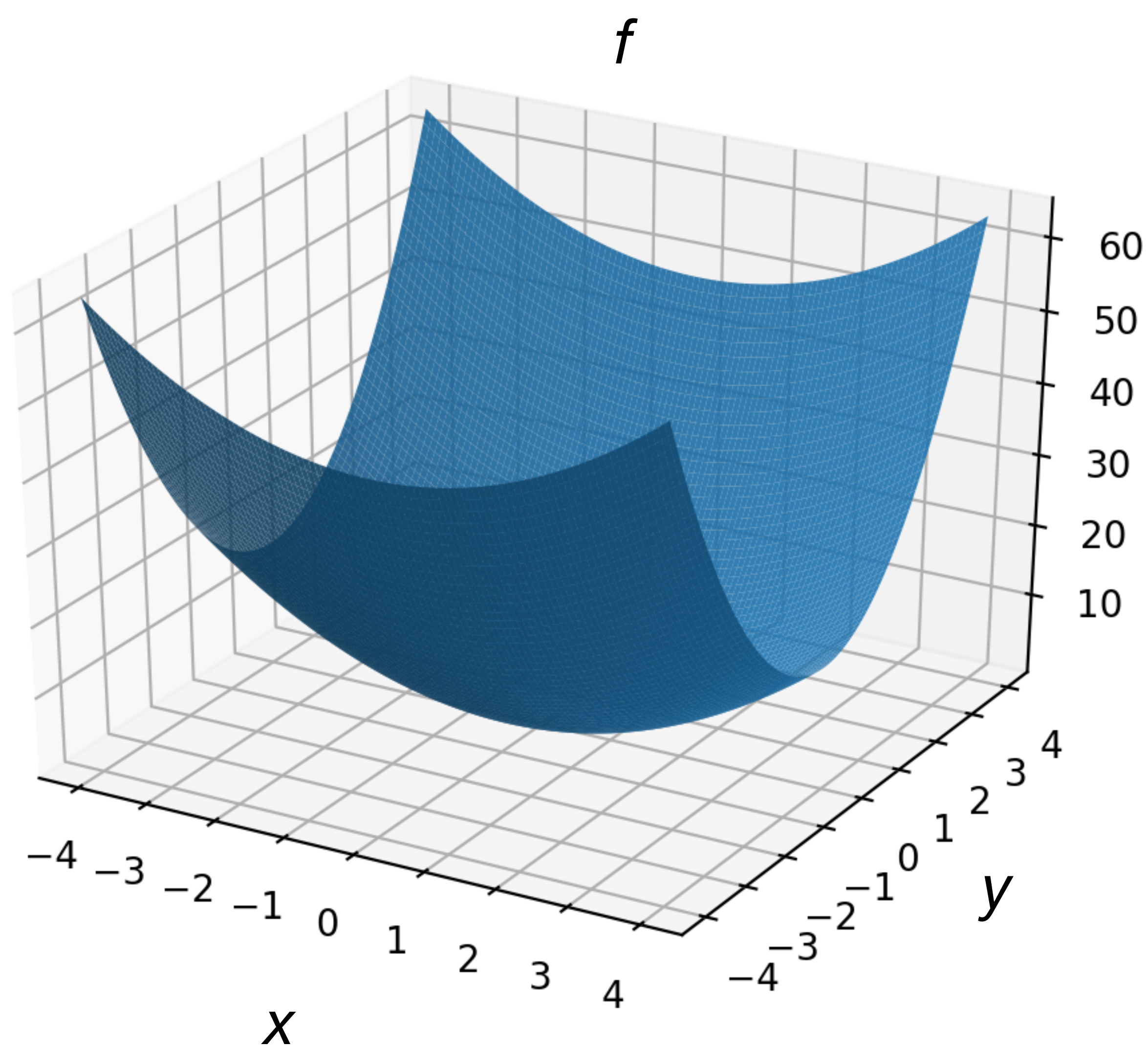
Lagrange multipliers

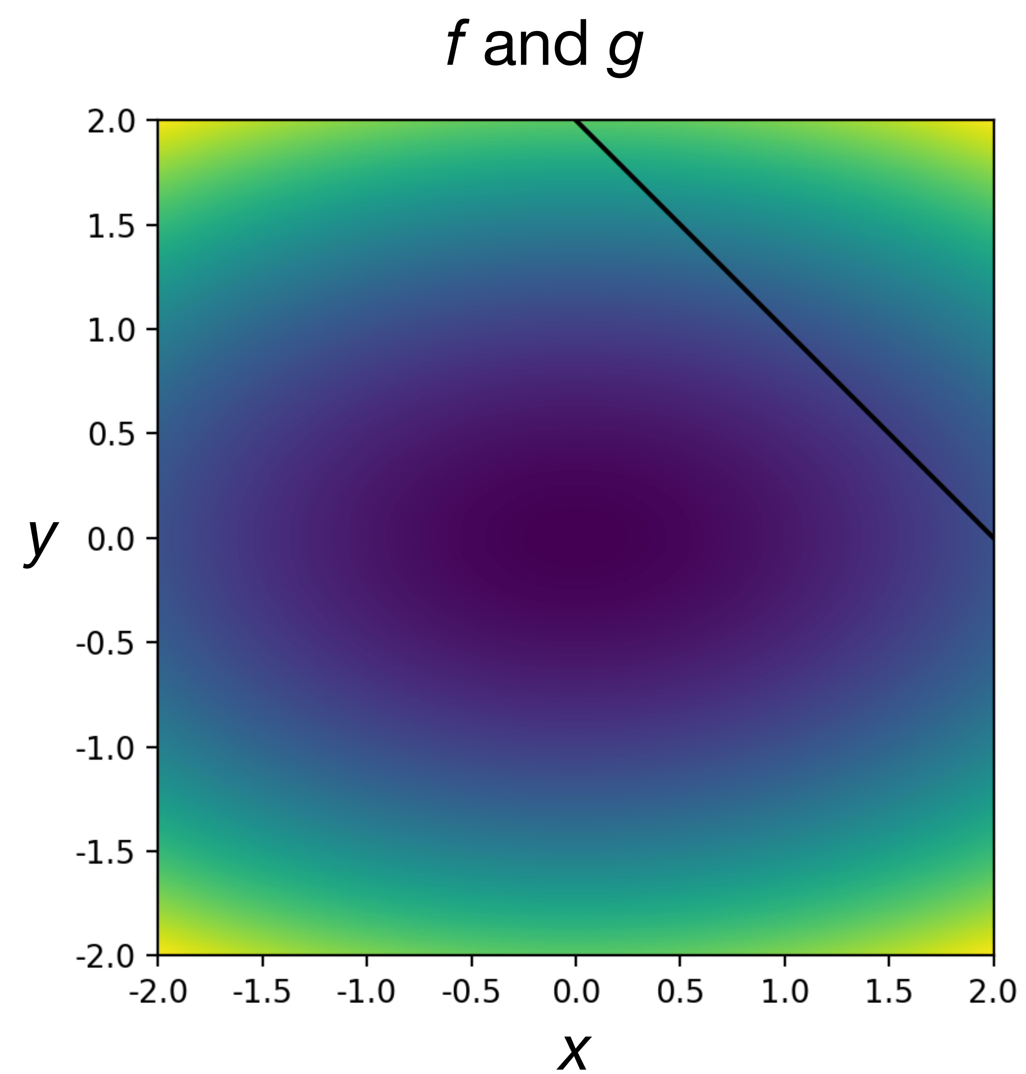
- The solution occurs at a critical point of *L*, i.e., where the derivative of *L* with respect to *x*, *y*, and *α* = 0.

$$\begin{aligned} L(x, y, \alpha) &= f(x, y) + \alpha g(x, y) \\ \frac{\partial L}{\partial x} &= 0 \\ \frac{\partial L}{\partial y} &= 0 \\ \frac{\partial L}{\partial \alpha} &= 0 \end{aligned}$$

Example

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





Example

$$\begin{aligned} f(x,y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\ L(x,y,\alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \end{aligned}$$

Example

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

Example

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

Example

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

Example

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

Example

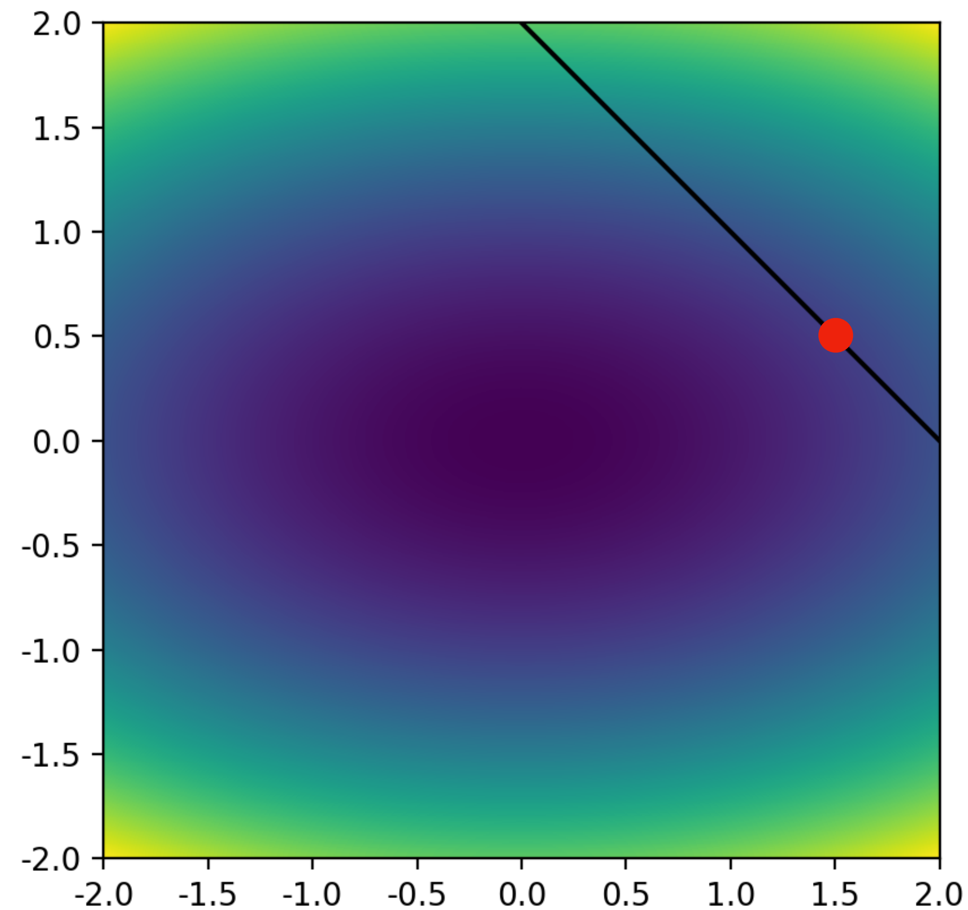
$$
\begin{aligned}
f(x,y) &= x^2+3y^2 \quad \text{subject to} \quad x+y=2 \\
L(x,y,\alpha) &= x^2+3y^2+\alpha(x+y-2) \\
\frac{\partial L}{\partial x} &= 2x+\alpha=0 \\
\frac{\partial L}{\partial y} &= 6y+\alpha=0 \\
\frac{\partial L}{\partial \alpha} &= x+y-2=0 \\
2x &= 6y \\
x &= 3y \\
3y+y-2 &= 0 \\
4y &= 2
\end{aligned}
$$

Example

$$
\begin{aligned}
f(x,y) &= x^2+3y^2 \quad \text{subject to} \quad x+y=2 \\
L(x,y,\alpha) &= x^2+3y^2+\alpha(x+y-2) \\
\frac{\partial L}{\partial x} &= 2x+\alpha=0 \\
\frac{\partial L}{\partial y} &= 6y+\alpha=0 \\
\frac{\partial L}{\partial \alpha} &= x+y-2=0 \\
2x &= 6y \\
x &= 3y \\
3y+y-2 &= 0 \\
4y &= 2 \\
y &= 1/2
\end{aligned}
$$

Example

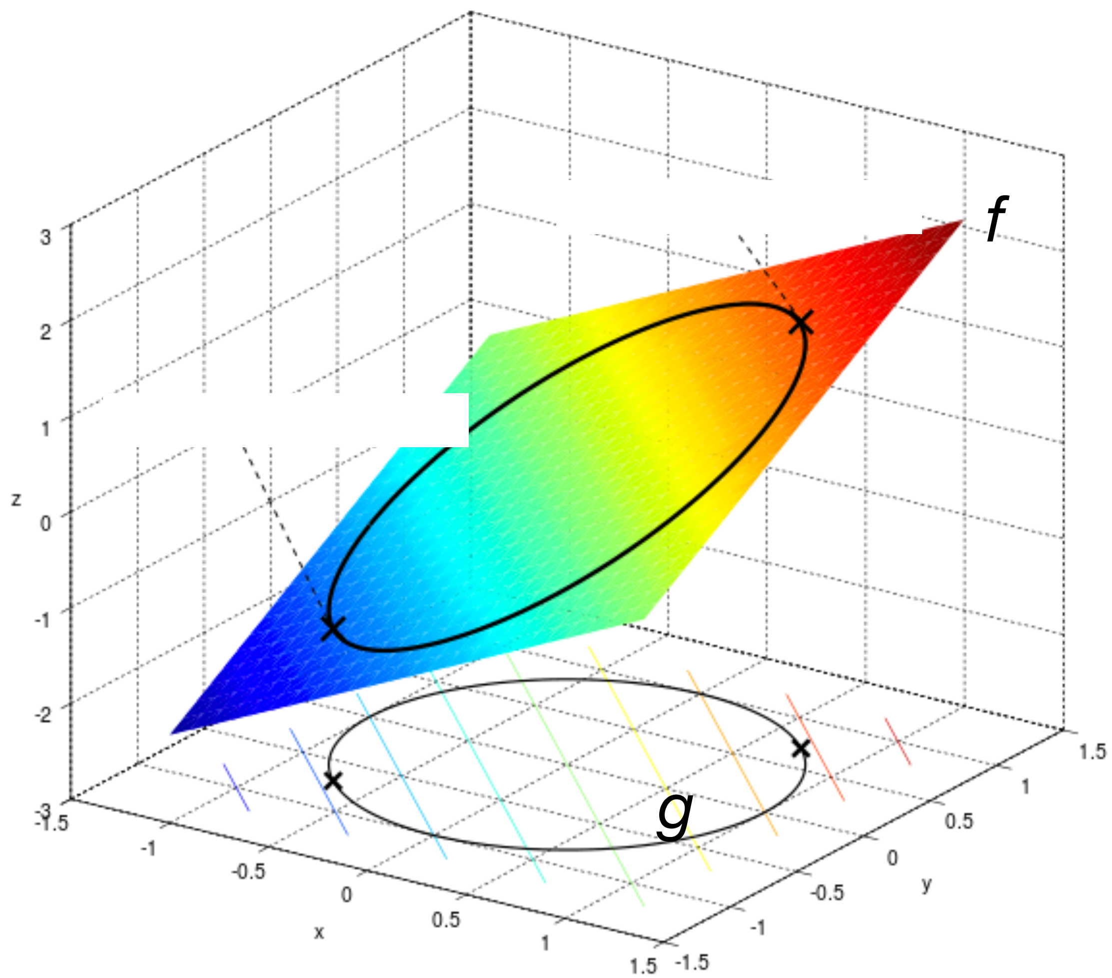
$$
\begin{aligned}
f(x,y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\
L(x,y,\alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \\
\frac{\partial L}{\partial x} &= 2x + \alpha = 0 \\
\frac{\partial L}{\partial y} &= 6y + \alpha = 0 \\
\frac{\partial L}{\partial \alpha} &= x + y - 2 = 0 \\
2x &= 6y \\
x &= 3y \\
3y + y - 2 &= 0 \\
4y &= 2 \\
y &= 1/2 \\
x &= 3/2
\end{aligned}
$$



Exercise

- Minimize:

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



Exercise

- Minimize:

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

Exercise

- Minimize:

$$\begin{aligned} f(x, y) &= x + y \quad \text{subject to} \quad x^2 + y^2 = 1 \\ L(x, y, \alpha) &= x + y + \alpha(x^2 + y^2 - 1) \end{aligned}$$

Exercise

- Minimize:

$$
\begin{aligned}
f(x,y) &= x+y \quad \text{subject to} \quad x^2+y^2=1 \\
L(x,y,\alpha) &= x+y+\alpha(x^2+y^2-1) \\
\frac{\partial L}{\partial x} &= 1+2\alpha x=0 \\
\frac{\partial L}{\partial y} &= 1+2\alpha y=0 \\
\frac{\partial L}{\partial \alpha} &= x^2+y^2-1=0
\end{aligned}
$$

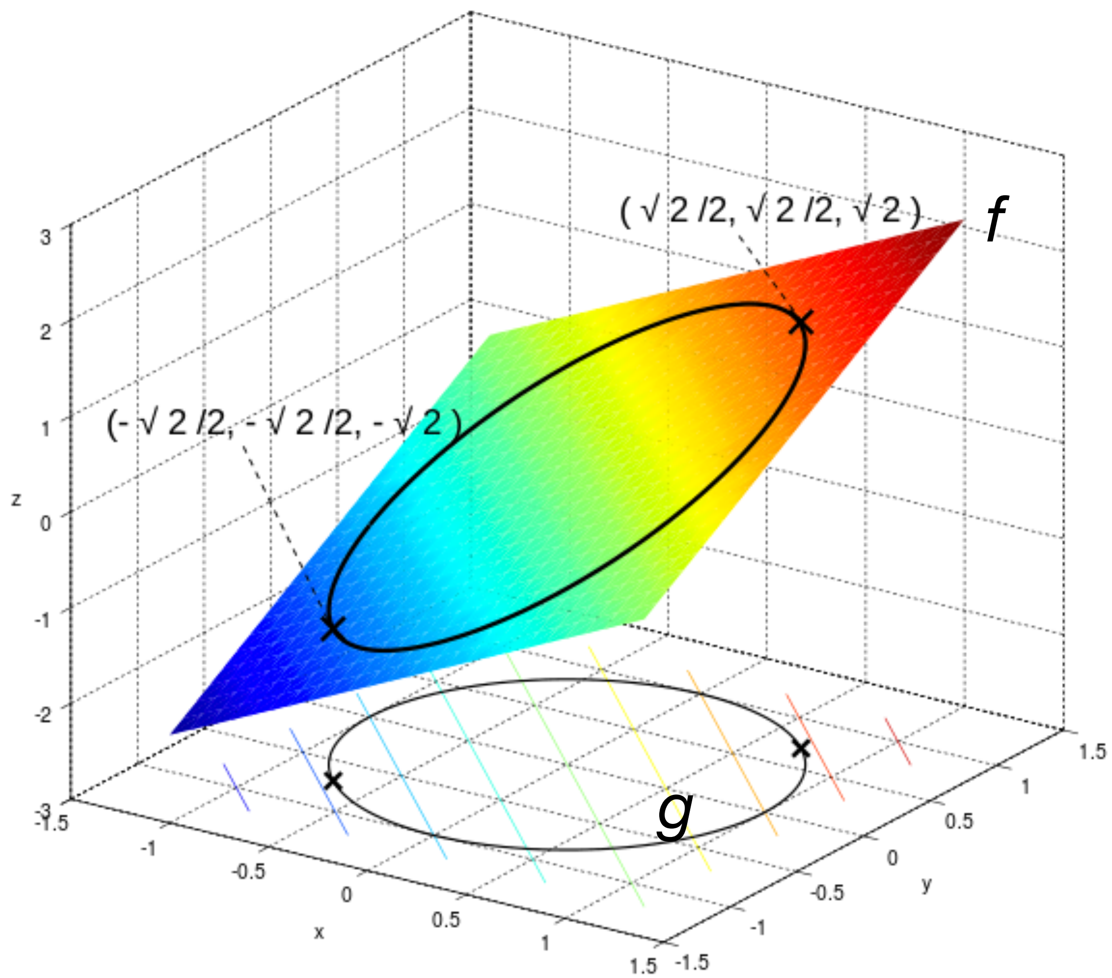
Exercise

- Minimize:

$$
\begin{aligned}
f(x,y) &= x + y \quad \text{subject to} \quad x^2 + y^2 = 1 \\
L(x,y,\alpha) &= x + y + \alpha(x^2 + y^2 - 1) \\
\frac{\partial L}{\partial x} &= 1 + 2\alpha x = 0 \\
\frac{\partial L}{\partial y} &= 1 + 2\alpha y = 0 \\
\frac{\partial L}{\partial \alpha} &= x^2 + y^2 - 1 = 0 \\
2\alpha x &= -1 \\
x &= -1/(2\alpha) \\
y &= -1/(2\alpha) = x \\
x^2 + (x)^2 - 1 &= 0 \\
2x^2 &= 1 \\
x^2 &= 1/2 \\
x &= y = \quad \pm 1/\sqrt{2}
\end{aligned}
$$

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

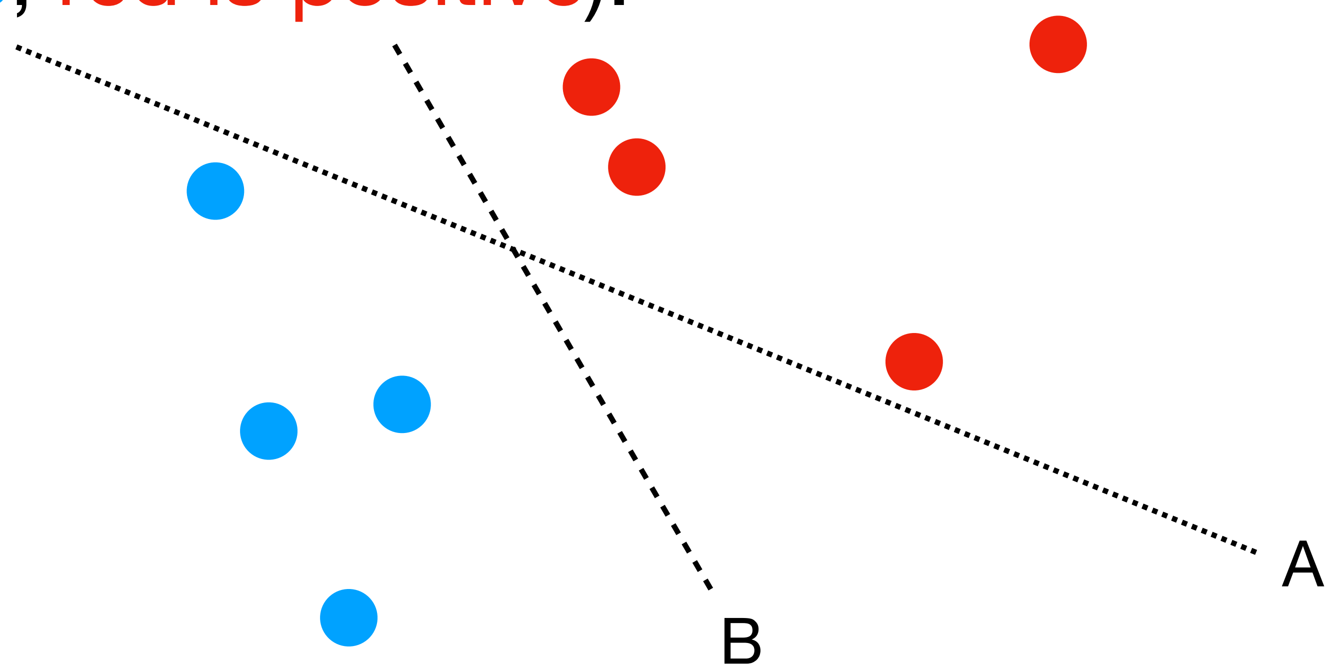
Support vector machines

- **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):

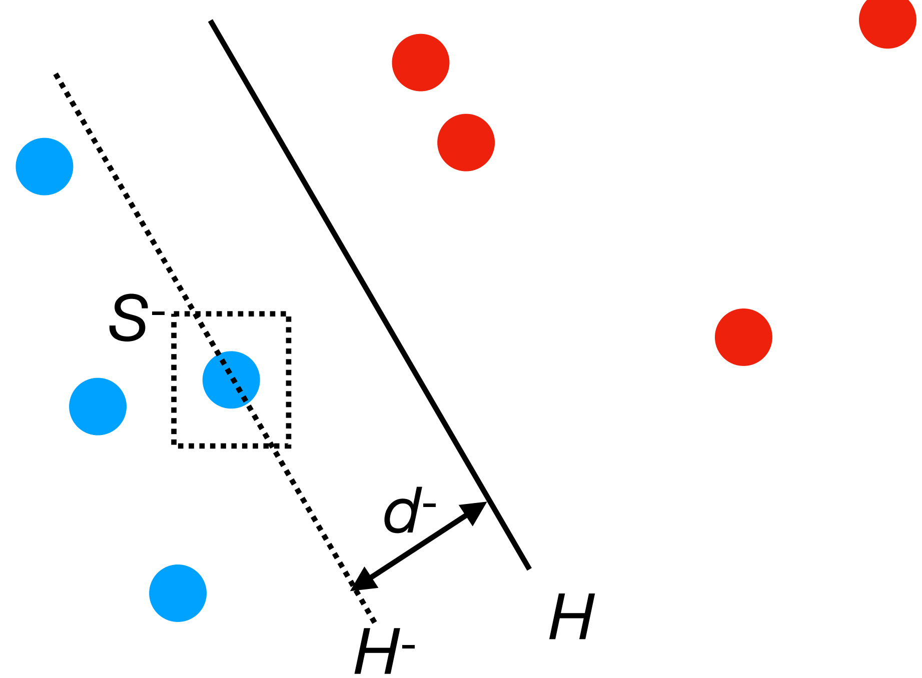


- Examples above the line will be classified as positive; examples below the line will be classified as negative.

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

Support vector machines

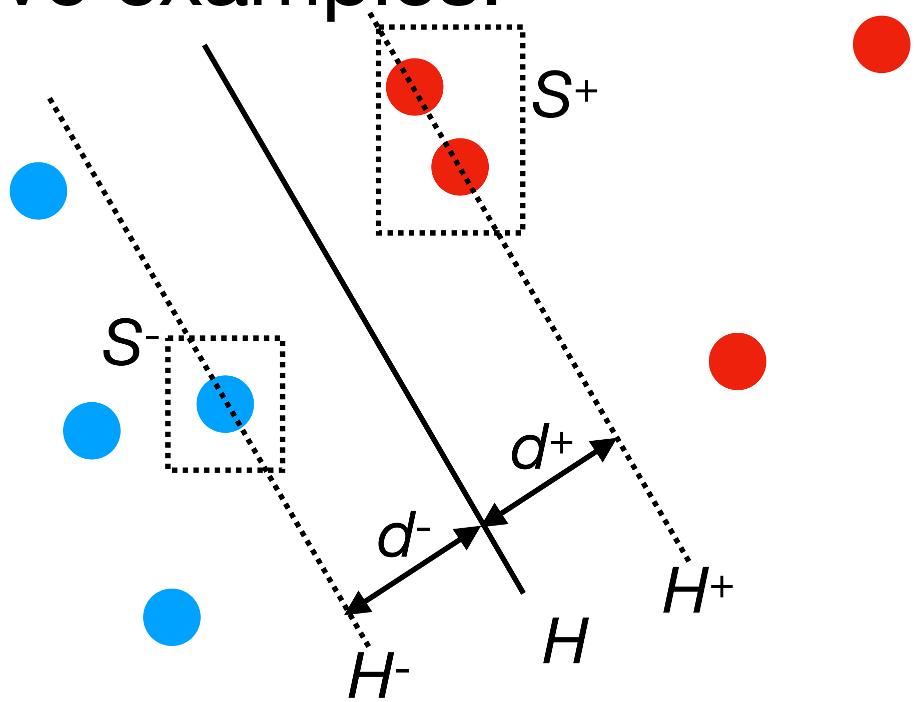
- For any hyperplane H that perfectly separates the positive from the negative examples:



- Find the subset S^- of - 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

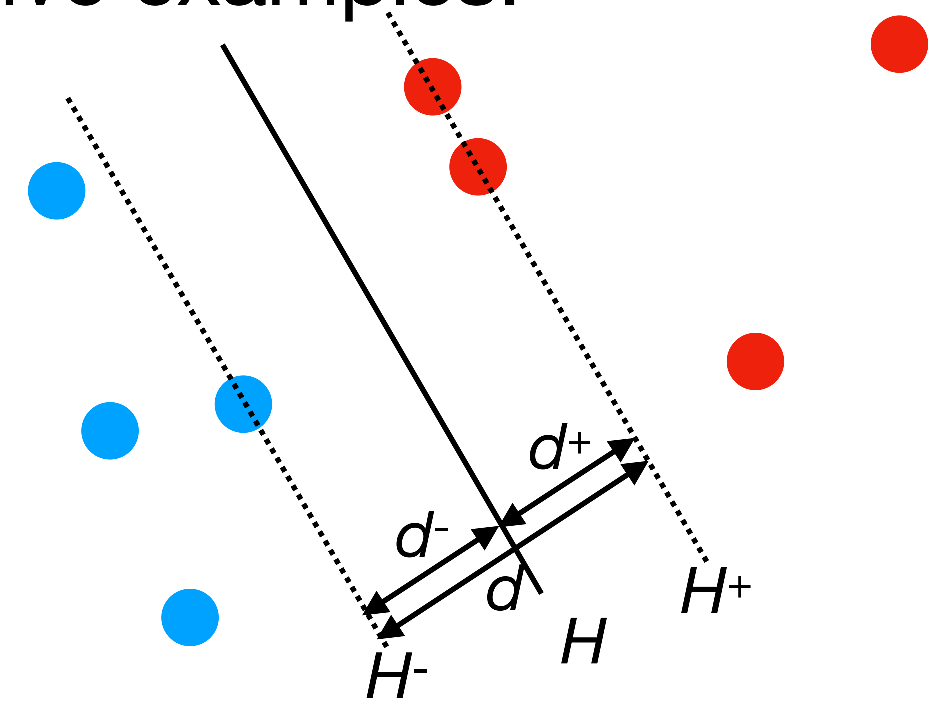
- For any hyperplane *H* that perfectly separates the positive from the negative examples:



  - 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^+.

Support vector machines

- 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.