EECS 553 Lecture 10
Machine Learning (ECE)
[class version]

Jeff Fessler

University of Michigan
©Laura Balzano and Jeff Fessler, UM, September 30, 2026, 15:54

F26

- PA3: **Fri. Nov. 6**: draft report.

Recap / Outline

© 2026 Fessler

- Recap
  - Logistic regression, softmax, cross entropy, multiclass log. reg.,
  - imbalanced data; hyperplanes
- Today
  - Hyperplanes
    - Definition, distances
    - Separating hyperplanes
  - Max-margin hyperplanes
  - Optimal soft-margin hyperplane: Prelude
  - Constrained optimization
    - Lagrangian
    - Duality
    - Convex constrained optimization
    - KKT conditions

Hyperplanes: Definition

© 2026 Fessler

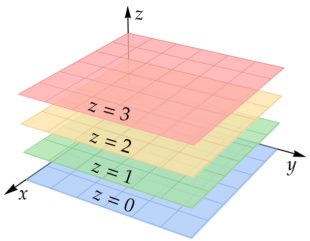


Figure 1: The xy-plane and several other horizontal planes.

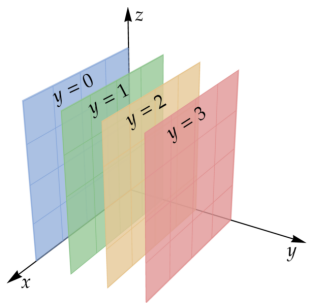


Figure 2: The xz-plane and several parallel planes.

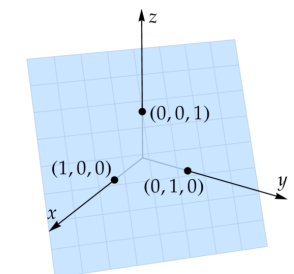


Figure 3: The plane $x + y + z = 1$.

Decision boundaries are often **hyperplanes** so we focus on those first.

- A **hyperplane** is a subset of $\mathbb{R}^d$ of the form

$$\mathcal{H} = \mathcal{H}(\boldsymbol{w}, b) = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{w}^\top \boldsymbol{x} + b = 0\},$$

  for some $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$.

- In general, a hyperplane in $\mathbb{R}^d$ is an **affine subspace** of **dimension** $d - 1$.

- The vector $\boldsymbol{w}$ is **orthogonal** to the hyperplane, and for this reason is called a **normal vector**.

Figure courtesy [2].

A **hyperplane** is a subset of $\mathbb{R}^d$ of the form

$$\mathcal{H} = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{w}^\top \boldsymbol{x} + b = 0\}, \quad \text{for some given } \boldsymbol{w} \in \mathbb{R}^d,\ b \in \mathbb{R}\ .$$

If $\boldsymbol{z} \in \mathcal{H}$, then $\boldsymbol{z} = t\boldsymbol{w} + \underbrace{(\boldsymbol{z} - t\boldsymbol{w})}_{\perp \text{ to } \boldsymbol{w}}$ for some $t \in \mathbb{R}$.

Hyperplanes: Unique representation?

© 2026 Fessler

For some given $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$, define the hyperplane $\mathcal{H} \triangleq \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{w}^\top \boldsymbol{x} + b = 0\}$.

Q10.1 Is the representation unique? If any $\tilde{\boldsymbol{w}} \in \mathbb{R}^d$ and $\tilde{b} \in \mathbb{R}$ represent the same hyperplane as $(\boldsymbol{w}, b)$, then $\tilde{\boldsymbol{w}} = \boldsymbol{w}$ and $\tilde{b} = b$.

A: True B: False **??**

- Distances to hyperplanes are crucial for classification problems.
- Find the distance from $\boldsymbol{z} \in \mathbb{R}^d$ to $\mathcal{H} \triangleq \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{w}^\top \boldsymbol{x} + b = 0\}$.
- The point closest to $\boldsymbol{z}$ on the **hyperplane** $\mathcal{H}$ can be written as $\hat{\boldsymbol{z}} = \boldsymbol{z} + t\boldsymbol{w}$ for a scalar t that satisfies

$$0 = \boldsymbol{w}^\top \hat{\boldsymbol{z}} + b = \boldsymbol{w}^\top (\boldsymbol{z} + t\boldsymbol{w}) + b \implies t = -\frac{\boldsymbol{w}^\top \boldsymbol{z} + b}{\boldsymbol{w}^\top \boldsymbol{w}}$$

- Thus the distance is

$$\|\hat{\boldsymbol{z}} - \boldsymbol{z}\|_2 = \|(\boldsymbol{z} + t\boldsymbol{w}) - \boldsymbol{z}\|_2 = |t| \, \|\boldsymbol{w}\|_2 = \frac{|\boldsymbol{w}^\top \boldsymbol{z} + b|}{\|\boldsymbol{w}\|_2^2} \|\boldsymbol{w}\|_2 = \frac{|\boldsymbol{w}^\top \boldsymbol{z} + b|}{\|\boldsymbol{w}\|_2}$$

Separating hyperplanes: Definition

© 2026 Fessler

- Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be training data with $\boldsymbol{x}_i \in \mathbb{R}^d$ for a **binary classification** problem.
- Assume $y_i \in \{-1, 1\}$.
- We say the training data are **linearly separable** iff there exist $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that

- In this case we call the following set a **separating hyperplane**:

$$\mathcal{H} \triangleq \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{w}^\top \boldsymbol{x} + b = 0\}.$$

Separating hyperplanes: Choices

© 2026 Fessler

Are all **separating hyperplanes** equally good?

Training data

Max-margin hyperplanes: Definition

© 2026 Fessler

- The **margin** ρ of a **separating hyperplane** is the distance from the hyperplane to the nearest training point:

$$\rho(\boldsymbol{w}, b) \triangleq \min_{i=1,\ldots,n} d(\mathcal{H}(\boldsymbol{w}, b), \boldsymbol{x}_i) = \min_{i=1,\ldots,n}$$

- Here we really mean min, not argmin!
- And here $d(\cdot, \cdot)$ is a distance function, not dimension d.

- An **optimal separating hyperplane** or **maximum-margin hyperplane** uses the solution of

$$\hat{\boldsymbol{w}}, \hat{b} = \arg\max_{\boldsymbol{w}, b} \rho(\boldsymbol{w}, b) \text{ such that}$$

Max-margin hyperplane: Example

© 2026 Fessler

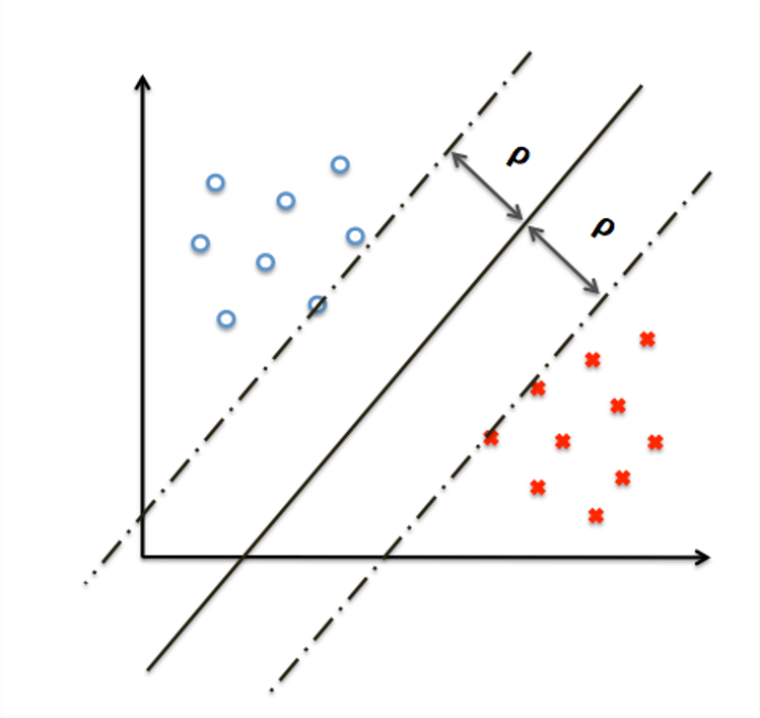


Figure courtesy Clay Scott.

Q10.2 For $n > d$, the max-margin hyperplane $\mathcal{H}$ is always unique? (Is $\mathcal{H}$ unique? Not asking whether $(b, \boldsymbol{w})$ is unique.)

A: True B: False **??**

Separating hyperplane: Canonical form

© 2026 Fessler

- The separability condition $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0$ holds for $(\alpha \boldsymbol{w}, \alpha b)$ if $\alpha > 0$, so there is a kind of scale ambiguity.
- A **separating hyperplane** is said to be in **canonical form** [3, §5.4.1] iff $\boldsymbol{w}$ and b are such that

$$\begin{cases} \geq 1, & \forall i \\ = 1, & \text{for some } i. \end{cases}$$

- Every separating hyperplane can be expressed in canonical form.
- Suppose $\mathcal{H} = \{\boldsymbol{x} : \boldsymbol{w}_1^\top \boldsymbol{x} + b_1 = 0\}$ is a separating hyperplane (not necessarily in canonical form). Let

$$m \triangleq ,$$

and define

$$\boldsymbol{w}_2 = \frac{\boldsymbol{w}_1}{m}, \quad b_2 = \frac{b_1}{m}.$$

Then $\boldsymbol{w}_2$, b_2 express $\mathcal{H}$ in canonical form.

Canonical form hyperplanes

© 2026 Fessler

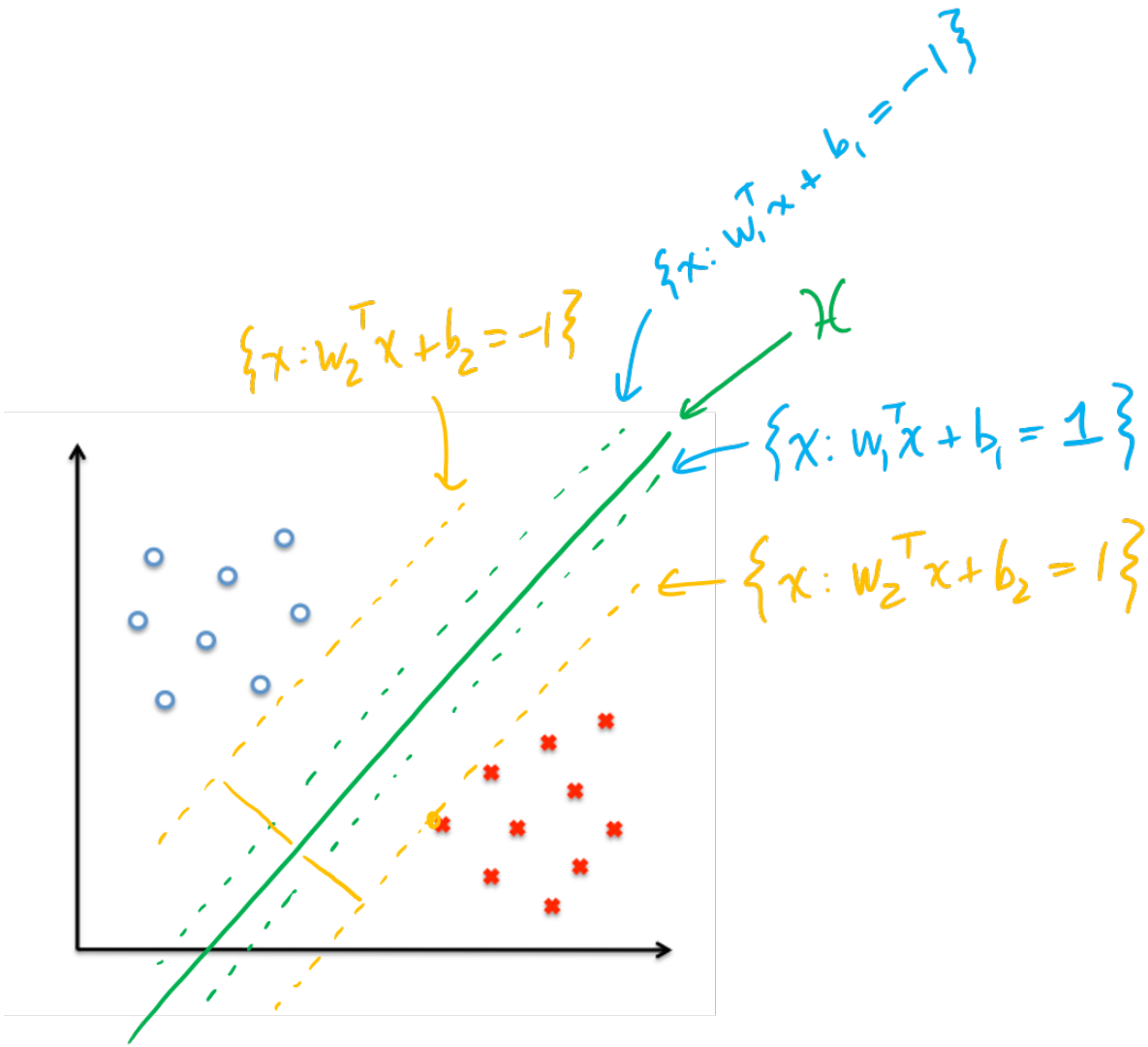


Figure courtesy Clay Scott.

- Thus we can define the max-margin hyperplane using

$$\begin{aligned} \arg\max_{\boldsymbol{w},b} \quad & \min_{i=1,\dots,n} \frac{|\boldsymbol{w}^\top \boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|_2} \\ \text{s.t. } \forall i \quad & y_i \left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1 \\ \exists i \quad & y_i \left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) = 1. \end{aligned}$$

- Equivalently:
- This is an example of a **quadratic program.**
- Proof. For some i:

Non-separable data

© 2026 Fessler

What if the data are not separable?

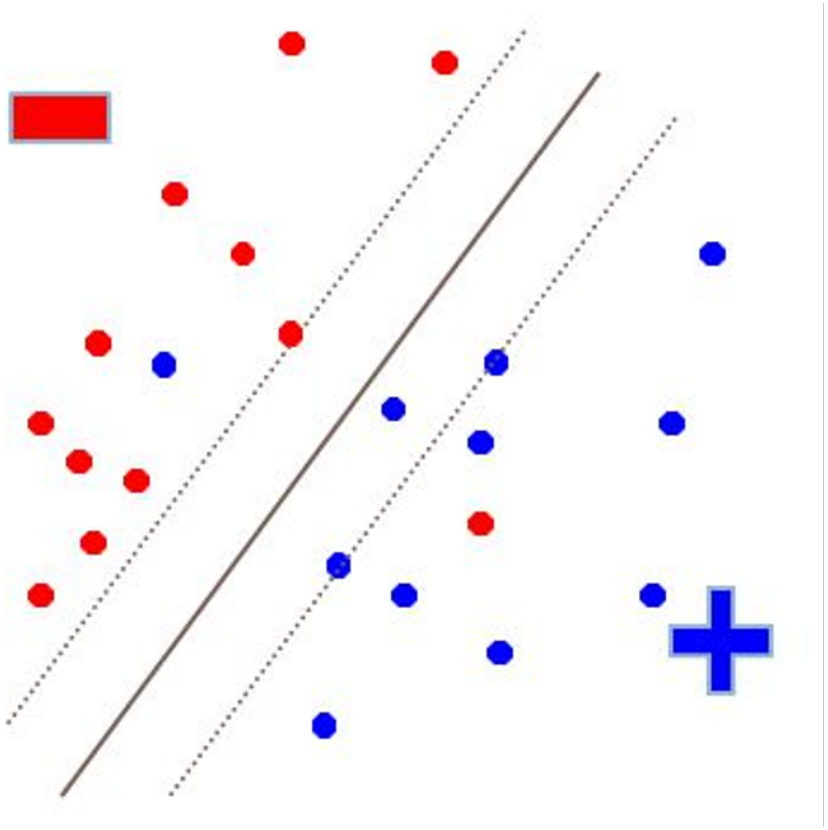


Figure courtesy Clay Scott.

Optimal soft-margin hyperplane: Prelude

© 2026 Fessler

- Introduce nonnegative **slack variables** $\xi_1, \dots, \xi_n \geq 0$.
  Greek letter ξ is pronounced "ksee."
- The optimal **soft-margin hyperplane** uses the solution of

$$\arg\min_{\boldsymbol{w},b} \min_{\boldsymbol{\xi}} \frac{1}{2} \|\boldsymbol{w}\|_2^2 + \qquad$$

$$\text{s.t. } y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq \qquad \forall i = 1, \dots, n$$

$$\xi_i \geq 0 \qquad \forall i = 1, \dots, n$$

- C is a user-defined parameter
- This is another **quadratic program**.
- The OSM hyperplane is a special case of a **support vector machine** (**SVM**).
- How do we solve such *constrained* optimization problems?

So far we have considered only convex optimization problems that are *unconstrained*, where the variable $\boldsymbol{u} \in \mathbb{R}^p$ can take any values in the p-dimensional space, where p denotes the number of model parameters.

The theory of **constrained optimization** allows us to consider problems where we must have $\boldsymbol{u} \in \mathcal{C} \subset \mathbb{R}^p$, like max margin or soft margin separating hyperplanes.

Outline:

- Lagrangian
- Duality
- KKT conditions

- A **constrained optimization** problem has the form

$$\begin{aligned} \underset{\boldsymbol{u} \in \mathbb{R}^p}{\arg\min} \quad & \\ \text{s.t.} \quad & \\ & \end{aligned}$$

where $\boldsymbol{u} \in \mathbb{R}^p$ and

- A point $\boldsymbol{u} \in \mathbb{R}^p$ is called **feasible** iff it satisfies all $r + s$ constraints above.

Lagrangian: Definition

© 2026 Fessler

- To solve such problems we usually use the **Lagrangian** function:

$$L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \triangleq$$

- $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_r]^\top \in \mathbb{R}^r$ and $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_s]^\top \in \mathbb{R}^s$ are called **Lagrange multipliers** or **dual variables**.

Lagrange dual function

© 2026 Fessler

- The **Lagrange dual function** is

$$L_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \triangleq$$

  - Here we really do mean min.
  - This is an *unconstrained* optimization problem in $\boldsymbol{u}$!
- The **dual optimization problem** is

$$\underset{\boldsymbol{\lambda}, \boldsymbol{\nu} \,:\, \lambda_i \geq 0}{\arg\max} \; L_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu}).$$

  - Note that $\boldsymbol{\lambda}$ must be nonnegative, but $\boldsymbol{\nu}$ is unconstrained.
  - Nonnegativity is simpler than constraints like $g_i(\boldsymbol{u}) \leq 0$.
- This is a **max-min optimization** problem.
- One reason this formulation is helpful is that L_{D} is a **concave function**.

L_{D} is a **concave function**, being the point wise minimum of a family of affine functions of $(\boldsymbol{\lambda}, \boldsymbol{\nu})$.

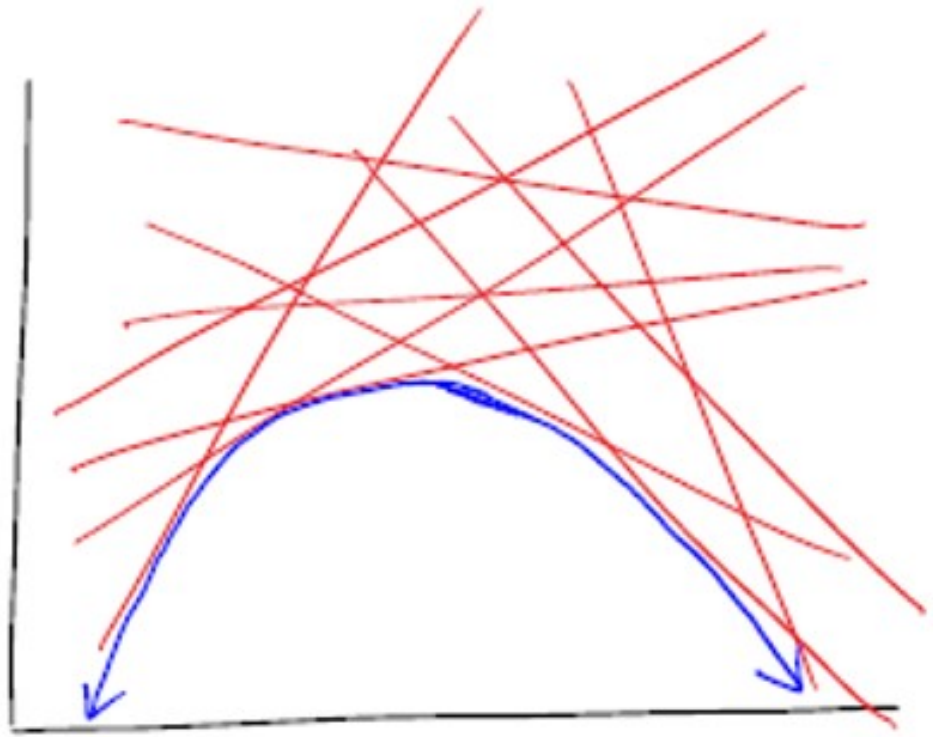


Illustration of $L_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu})$, courtesy Clay Scott. Horizontal axis represents $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ in 1D.

Similarly, the **Lagrange primal function** is

$$L_{\mathrm{P}}(\boldsymbol{u}) \triangleq \max_{\boldsymbol{\lambda},\boldsymbol{\nu}\,:\,\lambda_i \geq 0} L(\boldsymbol{u},\boldsymbol{\lambda},\boldsymbol{\nu})$$

$$= \max_{\boldsymbol{\lambda},\boldsymbol{\nu}\,:\,\lambda_i \geq 0} \left(f(\boldsymbol{u}) + \sum_{i=1}^{r} \lambda_i g_i(\boldsymbol{u}) + \sum_{j=1}^{s} \nu_j h_j(\boldsymbol{u}) \right).$$

Q10.3 Which of the following is true? (This uses only what we have learned so far.)

A: $L_{\mathrm{P}}(\boldsymbol{u}) \leq L_{\mathrm{D}}(\boldsymbol{\lambda},\boldsymbol{\nu})$

B: $L_{\mathrm{P}}(\boldsymbol{u}) = f(\boldsymbol{u}) + \sum_{i=1}^{r} \lambda_i g_i(\boldsymbol{u})$ if $\boldsymbol{u}$ is **feasible**.

C: $L_{\mathrm{P}}(\boldsymbol{u}) = f(\boldsymbol{u})$ if $\boldsymbol{u}$ is feasible.

D: none of these

Duality I

Denote the optimal cost function values of the **primal problem** and **dual problem**, respectively, by

$$p^* \triangleq \min_{\boldsymbol{u}} L_{\mathrm{P}}(\boldsymbol{u}) = \min_{\boldsymbol{u}} \max_{\boldsymbol{\lambda},\boldsymbol{\nu}\,:\,\lambda_i \geq 0} L(\boldsymbol{u},\boldsymbol{\lambda},\boldsymbol{\nu})$$

$$d^* \triangleq \max_{\boldsymbol{\lambda},\boldsymbol{\nu}\,:\,\lambda_i \geq 0} L_{\mathrm{D}}(\boldsymbol{\lambda},\boldsymbol{\nu}) = \max_{\boldsymbol{\lambda},\boldsymbol{\nu}\,:\,\lambda_i \geq 0} \min_{\boldsymbol{u}} L(\boldsymbol{u},\boldsymbol{\lambda},\boldsymbol{\nu}).$$

These values are related by the following property (Theorem) called **weak duality**:

- The difference is called the **duality gap**.
- If $p^* = d^*$, we say **strong duality** holds.

Proof. Let $\tilde{\boldsymbol{u}}$ be feasible. Then for any $\boldsymbol{\lambda}, \boldsymbol{\nu}$ with $\lambda_i \geq 0$:

$$L(\tilde{\boldsymbol{u}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\tilde{\boldsymbol{u}}) + \sum_{i=1}^{r} \lambda_i \underbrace{g_i(\tilde{\boldsymbol{u}})}_{\leq 0} + \sum_{j=1}^{s} \nu_j \underbrace{h_j(\tilde{\boldsymbol{u}})}_{= 0} \leq f(\tilde{\boldsymbol{u}}).$$

When we unconstrain $\boldsymbol{u}$, the cost must be no higher, so:

$$L_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{u} \in \mathbb{R}^p} L(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\tilde{\boldsymbol{u}}).$$ (When we unconstrain $\boldsymbol{u}$, cost must be lower.)

This inequality holds for any feasible $\tilde{\boldsymbol{u}}$, so

$$L_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \inf_{\tilde{\boldsymbol{u}} \text{ feasible}} f(\tilde{\boldsymbol{u}}) = p^*.$$

Taking the max over $\boldsymbol{\lambda}, \boldsymbol{\nu} : \lambda_i \geq 0$, we have

$$d^* = \sup_{\boldsymbol{\lambda}, \boldsymbol{\nu} : \lambda_i \geq 0} L_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*. \qquad \square$$

Constrainted optimization problem statement:

$$\begin{aligned} &\arg\min_u \frac{1}{2}u^2 \\ &\text{s.t. } (u-3)^2 \le 1 \end{aligned} \qquad f(u) = \frac{1}{2}u^2, \quad g(u) = (u-3)^2 - 1 \le 0$$

- Lagrangian: $L(u,\lambda) = f(u) + \lambda g(u) = \frac{1}{2}u^2 + \lambda\Big((u-3)^2 - 1\Big)$
  A parabola in u with positive curvature because $\lambda \ge 0$.
- Lagrange dual: $L_{\mathrm{D}}(\lambda) = \inf_u L(u,\lambda)$
  $0 = \frac{\partial}{\partial u} L(u,\lambda) = u + \lambda 2(u-3) \implies u_\lambda = \frac{6\lambda}{1+2\lambda}$
  $L_{\mathrm{D}}(\lambda) = L(u_\lambda,\lambda) = f(u_\lambda) + \lambda g(u_\lambda) = \frac{1}{2}u_\lambda^2 + \lambda\Big((u_\lambda-3)^2 - 1\Big) =$
  $\frac{1}{2}\left(\frac{6\lambda}{1+2\lambda}\right)^2 + \lambda\left(\left(\frac{6\lambda}{1+2\lambda} - 3\right)^2 - 1\right) = \cdots = \frac{8\lambda - 2\lambda^2}{1+2\lambda}.$
  $\implies \frac{\mathrm{d}}{\mathrm{d}\lambda} L_{\mathrm{D}}(\lambda) = -4\frac{\lambda^2+\lambda-2}{(1+2\lambda)^2} = 0$ when $\lambda \in \{\frac{-1\pm\sqrt{1+8}}{2}\} = \{-2, 1\}$.
- Maximize the dual (easy nonnegativity constraint):
  $\lambda_* = \arg\max_{\lambda\ge 0} L_{\mathrm{D}}(\lambda) = 1$ (positive stationary point)
- Primal solution: $u_* = u_{\lambda_*} = \frac{6\lambda_*}{1+2\lambda_*} = 2$ (as expected)
- Dual optimal cost value: $d^* = L_{\mathrm{D}}(\lambda_*) = \frac{8\lambda_* - 2\lambda_*^2}{1+2\lambda_*} = \frac{6}{3} = 2.$
- Primal optimal cost value: $p^* = f(\boldsymbol{u}_*) = f(2) = 2 = d^*.$

Lagrange Example: 1D figure

© 2026 Fessler



Figure by Gemini

$$\begin{array}{l} \arg\min\limits_{u} u^2 \\ \text{s.t. } u \geq 3 \end{array} \qquad f(u) = u^2, \quad g(u) = 3 - u \leq 0.$$

- Lagrangian: $L(u, \lambda) = f(u) + \lambda g(u) = u^2 + \lambda(3 - u)$
- Lagrange dual: $L_{\mathrm{D}}(\lambda) = \inf_u L(u, \lambda)$
  $0 = \frac{\partial}{\partial u} L(u, \lambda) = 2u - \lambda \implies u_\lambda = \lambda/2$
  $L_{\mathrm{D}}(\lambda) = L(u_\lambda, \lambda) = u_\lambda^2 + \lambda(3 - u_\lambda) = (\lambda/2)^2 + \lambda(3 - \lambda/2) = 3\lambda - \lambda^2/4$
- Maximize the dual (easy nonnegativity constraint):
  $\arg\max_{\lambda \geq 0} L_{\mathrm{D}}(\lambda) = \arg\max_{\lambda \geq 0} = 3\lambda - \lambda^2/4 \implies \lambda_* = 6$
- Primal solution: $u_* = u_{\lambda_*} = \lambda_*/2 = 3$ (as expected).
- Dual optimal cost value: $d^* = L_{\mathrm{D}}(\lambda_*) = 3\lambda_* - \lambda_*^2/4 = 9$
- Primal optimal cost value: $p^* = f(u_*) = f(3) = 3^2 = 9 = d^*$

Recall general form of **constrained optimization**:

$$\begin{aligned} \underset{\boldsymbol{u} \in \mathbb{R}^p}{\arg\min} \; & f(\boldsymbol{u}) \\ \text{s.t.} \quad & g_i(\boldsymbol{u}) \le 0, \quad i = 1, \dots, r \\ & h_j(\boldsymbol{u}) = 0, \quad j = 1, \dots, s \end{aligned}$$

where $f, g_1, \dots, g_r, h_1, \dots, h_s : \mathbb{R}^p \mapsto \mathbb{R}$.

This constrained optimization problem is called **convex** iff

- f and $g_1, \dots, g_r$ are
- $h_1, \dots, h_s$ are

CQ: Unconstrained optimization refresh

© 2026 Fessler

Assume f is differentiable. Revisit necessary & sufficient conditions.

Q10.4 For **unconstrained optimization,** we know $\nabla f(\boldsymbol{u}^*) = \mathbf{0}$ is $\boxed{?}$ for $\boldsymbol{u}^*$ to be a global minimizer, and $\boxed{?}$.

A: necessary and sufficient / $\boldsymbol{u}^*$ is unique if f is additionally strictly convex

B: necessary / sufficient only if f is strictly convex

C: necessary / sufficient if f is convex

D: sufficient / necessary and sufficient if f is convex

E: none of these

The next two results generalize these properties to **constrained optimization**.

KKT conditions: Illustration

© 2026 Fessler

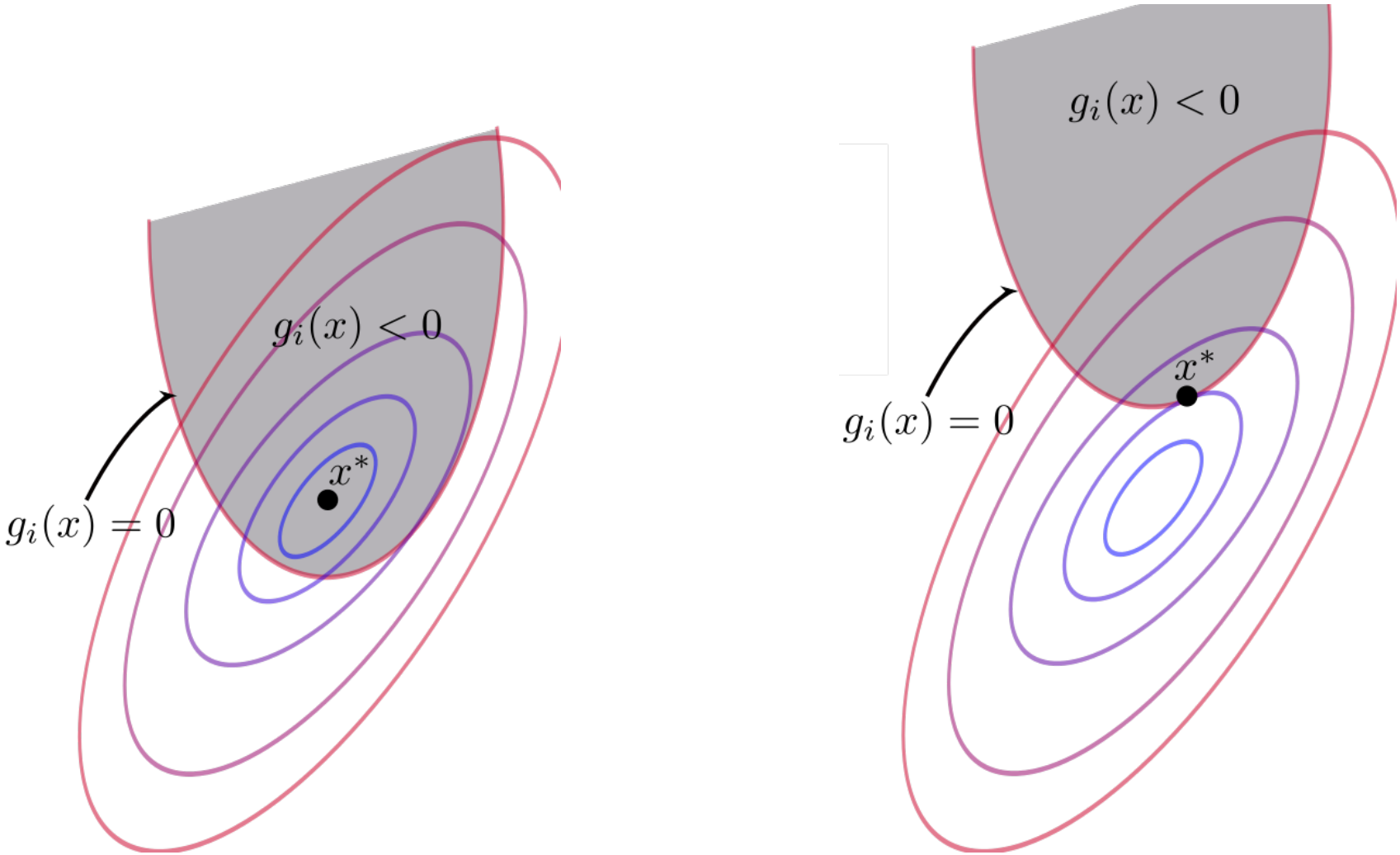


Figure from wikipedia [7].

Theorem (Necessity: Strong duality and optimality imply the KKT conditions hold [1, 8.5.2])

*If $p^* = d^*$ (strong duality), $\boldsymbol{u}^*$ is primal optimal, and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ are dual optimal, then the **Karesh-Kuhn-Tucker** (**KKT**) conditions hold:*

1. $\nabla_{\boldsymbol{u}} L(\boldsymbol{u}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = \nabla_{\boldsymbol{u}} f(\boldsymbol{u}^*) + \sum_{i=1}^{r} \lambda_i^* \nabla g_i(\boldsymbol{u}^*) + \sum_{j=1}^{s} \nu_j^* \nabla h_j(\boldsymbol{u}^*)$ (***stationarity*** *of the Lagrangian)*
2. $g_i(\boldsymbol{u}^*) \leq 0, \ \forall i$ (***primal feasibility***)
3. $h_j(\boldsymbol{u}^*) = 0, \ \forall j$ (“”)
4. $\lambda_i^* \geq 0, \ \forall i$ (***dual feasibility***)
5. (***complementary slackness***).

Proof uses the **hyperplane separation theorem**.

CQ: KKT necessary conditions

© 2026 Fessler

Q10.5 What is the dimension of the $\mathbf{0}$ above?
Enter one of $d, i, j, m, n, p, q, r, s$ with no spaces!

Theorem (**Sufficiency**: For **convex** problems, the KKT conditions imply both optimality and strong duality)

- **Constraint qualifications** provide guarantees that strong duality holds for certain convex problems, which in turn implies that a point is optimal See [8, §5.5].
- The most common is **Slater's constraint qualification** for convex problems that states sufficient conditions in terms of the inequality constraints[1], namely that there exists a strictly feasible point, *i.e.*, strict inequalities: $g_i(\boldsymbol{u}) < 0$.

[1] If there are affine inequality constraints, they can be ignored for this condition.

KKT (in)sufficiency example [Read]

© 2026 Fessler

- Consider the (convex and differentiable) problem

$$\arg\min_u \; u$$
$$\text{s.t. } g(u) \triangleq u^2 \le 0.$$

- Clearly the optimal $u^* = 0$. (But Slater's condition does not hold.)
- The Lagrangian is $L(u, \lambda) = f(u) + \lambda g(u) = u + \lambda u^2$
- The Lagrange dual is

$$L_{\mathrm{D}}(\lambda) = \min_u L(u, \lambda) = \begin{cases} -1/(4\lambda), & \lambda > 0 \text{ for } u = -1/(2\lambda) \\ -\infty, & \lambda \le 0 \end{cases}$$

- The dual problem is $\arg\max_{\lambda \ge 0} L_{\mathrm{D}}(\lambda) = \arg\max_{\lambda \ge 0} -1/(4\lambda)$ never reached for finite λ
- Stater's condition fails
- Check first KKT condition (stationarity):
$\frac{\partial f}{\partial u}(u^*) + \lambda^* \frac{\partial g}{\partial u}(u^*) = 1 + \lambda^*(2u^*) = 1 + \lambda^* \cdot 0 \stackrel{?}{=} 0$
This equality cannot be satisfied for any λ^*.

Usefulness of KKT

© 2026 Fessler

- Sometimes it is easier (analytically or computationally) to solve the dual, and then use the dual solution to solve the primal problem.
- For example: suppose strong duality holds. If $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is dual optimal, then (by the KKT necessity theorem) any primal optimal point $\boldsymbol{u}^*$ is a solution of

- We can use this equality to find a $\boldsymbol{u}^*$. In other words, we can recover a primal solution from a dual solution.

Certificates [Read]

© 2026 Fessler

Recall that $L_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq d^* \leq p^* \leq L_{\mathrm{P}}(\boldsymbol{u})$.

- If an algorithm finds feasible $\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu}$, then the gap $L_{\mathrm{P}}(\boldsymbol{u}) - L_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is called a **certificate of suboptimality**.
- When **strong duality** holds, there exist *arbitrarily good certificates*, *i.e.*, the gap can be made arbitrarily small.
- For convex problems, finding variables $(\boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ that satisfy the KKT conditions provides a **certificate of optimality**.
  This is sometimes also called a **dual certificate**.
- Take-away. First check if strong duality holds, or else the effort to find KKT points may be fruitless!

Examples include [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20].

References I

[1] K. P. Murphy.
Probabilistic machine learning: An introduction.
MIT Press, 2022.

[2] Jim Belk.
8.2 planes and hyperplanes, 2021.
https://e.math.cornell.edu/people/belk/writing/LinearODEs8.pdf.

[3] V. N. Vapnik.
The nature of statistical learning theory.
Springer, New York, 1999.

[4] C. Bishop.
Pattern recognition and machine learning.
Springer, 2006.

[5] T. Serafini, G. Zanghirati, and L. Zanni.
Gradient projection methods for quadratic programs and applications in training support vector machines.
Optimization Methods and Software, 20(2-3):353–78, 2005.

[6] R. O. Duda, P. E. Hart, and D. G. Stork.
Pattern classification.
Wiley, New York, 2001.

[7] Onmyphd.
Karush–kuhn–tucker conditions inequality constraint diagram, 2013.
http://www.onmyphd.com/?p=kkt.karush.kuhn.tucker, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=27331272.

[8] Stephen Boyd and Lieven Vandenberghe.
Convex Optimization.
Cambridge University Press, 2004.

References II

[9] Emmanuel Candes and Benjamin Recht.
Exact matrix completion via convex optimization.
Communications of the ACM, 55(6):111–119, 2012.

[10] David Gross.
Recovering low-rank matrices from few coefficients in any basis.
IEEE Transactions on Information Theory, 57(3):1548–1566, 2011.

[11] Ali Jalali, Yudong Chen, Sujay Sanghavi, and Huan Xu.
Clustering partially observed graphs via convex optimization.
In *ICML*, 2011.

[12] Paul Hand.
Conditions for dual certificate existence in semidefinite rank-1 matrix recovery.
In *Proceedings of the 10th International Conference on Sampling Theory and Applications*, 2013.

[13] Xiaodong Li and Vladislav Voroninski.
Sparse signal recovery from quadratic measurements via convex programming.
SIAM Journal on Mathematical Analysis, 45(5):3019–3033, 2013.

[14] Qiuwei Li, Ashley Prater, Lixin Shen, and Gongguo Tang.
Overcomplete tensor decomposition via convex optimization.
In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 53–56. IEEE, 2015.

[15] Stanley Osher and Wotao Yin.
Sparse recovery via ℓ_1 and l_1 optimization.
Notices of the International Congress of Chinese Mathematicians, 3(1):4–10, 2015.

[16] Carlos Fernandez-Granda.
Super-resolution of point sources via convex programming.
Information and Inference: A Journal of the IMA, 5(3):251–303, 2016.

References III

[17] Takayuki Iguchi, Dustin G Mixon, Jesse Peterson, and Soledad Villar.
Probably certifiably correct k-means clustering.
Mathematical Programming, 165(2):605–642, 2017.

[18] Chong You, Zhihui Zhu, Qing Qu, and Yi Ma.
Robust recovery via implicit bias of discrepant learning rates for double over-parameterization.
Advances in Neural Information Processing Systems, 33:17733–17744, 2020.

[19] Shunta Akiyama and Taiji Suzuki.
On learnability via gradient method for two-layer relu neural networks in teacher-student setting.
In *International Conference on Machine Learning*, pages 152–162. PMLR, 2021.

[20] Kyle Gilman, Sam Burer, and Laura Balzano.
A semidefinite relaxation for sums of heterogeneous quadratics on the stiefel manifold.
arXiv preprint arXiv:2205.13653, 2022.