CS5339 Lecture Notes #6a:
Dual SVM & A Detour Into Convex Analysis

Jonathan Scarlett

March 31, 2021

Useful references:

- Blog posts by Jeremy Kun on Lagrange multipliers,[1] duality for linear programming,[2] and duality for the support vector machine (SVM)[3]
- Blog posts by Sébastien Bubeck on Lagrangian duality[4] and SVM+duality+kernels[5]
- Part I of Boyd and Vandenberghe's "Convex Optimization" book[6]
- Boyd's lectures on convex optimization, available on YouTube
- Supplementary notes lec8a.pdf
- Section 12.1 of "Understanding Machine Learning" book

1 Convex Sets and Functions

Basic definitions.

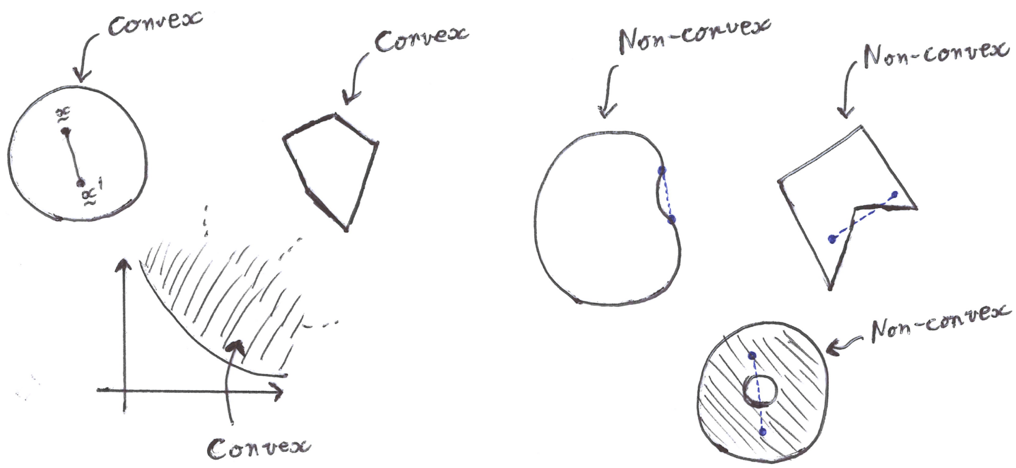
- A set D (e.g., a subset of $\mathbb{R}^d$) is said to be a *convex set* if, for all $\mathbf{x} \in D$ and $\mathbf{x}' \in D$, it holds that

$$\lambda \mathbf{x} + (1-\lambda)\mathbf{x}' \in D$$

  for all $\lambda \in [0,1]$

  - In words (roughly): Draw a straight line between any two points in D. This whole line segment must also lie within D.
  - Examples:

[1] `http://jeremykun.com/2013/11/30/lagrangians-for-the-amnesiac/`
[2] `http://jeremykun.com/2014/06/02/linear-programming-and-the-most-affordable-healthy-diet-part-1/`
[3] `http://jeremykun.com/2012/12/09/neural-networks-and-backpropagation/`
[4] `http://blogs.princeton.edu/imabandit/2013/02/21/orf523-lagrangian-duality/`
[5] `http://blogs.princeton.edu/imabandit/2013/02/26/orf523-classification-svm-kernel-learning/`
[6] `http://web.stanford.edu/~boyd/cvxbook/`

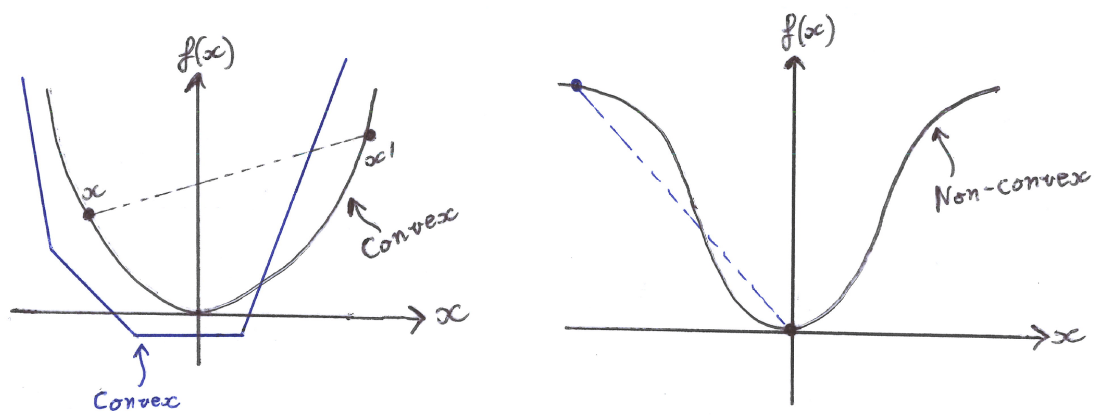


- A function $f : D \to \mathbb{R}$ is said to be a *convex function* if, for all $\mathbf{x} \in D$ and $\mathbf{x}' \in D$, it holds that

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{x}') \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{x}')$$

  for all $\lambda \in [0, 1]$. Implicitly, this requires that the domain D is a convex set.

  - In words (roughly): Draw a straight line between $(\mathbf{x}, f(\mathbf{x}'))$ and $(\mathbf{x}', f(\mathbf{x}'))$. For inputs in between $\mathbf{x}$ and $\mathbf{x}'$, the function lies below this straight line.
  - Illustration:



  - We say that $f(\mathbf{x})$ is a *concave function* if $-f(\mathbf{x})$ is a convex function.
  - Convex = “bowl-shaped” ($\cup$), concave = “arch-shaped” ($\cap$)
  - A function is simultaneously convex and concave $\iff$ it is affine (i.e., a “straight line” (or plane)).
  - Key property. For a convex function, any local minimum is also a global minimum.

Other examples.

- Convex functions: $\|\mathbf{x}\|^2$, e^x, e^{-x}, $\log \sum_{i=1}^d e^{x_i}$, and many more.
- Concave functions: $-\|\mathbf{x}\|^2$, $\log x$, $\log \det \mathbf{X}$, $\sum_{i=1}^d x_i \log \frac{1}{x_i}$, and many more.

Equivalent definitions of convexity.

- Recall the notions of gradient and Hessian for $\mathbf{x} = [x_1, \dots, x_d]^T$:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}, \qquad \nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}.$$

- (First order) If f is differentiable, then it is convex if and only if

$$f(\mathbf{x}') \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{x}' - \mathbf{x})$$

  for all $\mathbf{x}, \mathbf{x}'$. (The function lies above its tangent plane)

- (Second order) If f is twice differentiable, then it is convex if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$$

  for all $\mathbf{x} \in D$. (The Hessian is positive semi-definite)

Operations that preserve convexity.

- If $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are convex, and α_1 and α_2 are positive, then $f(\mathbf{x}) = \alpha_1 f_1(\mathbf{x}) + \alpha_1 f_2(\mathbf{x})$ is convex. By induction, a similar statement holds for $\sum_{\ell=1}^{L} \alpha_\ell f_\ell(\mathbf{x})$ also for $L > 2$.
- If $f_1(\mathbf{x}), \dots, f_L(\mathbf{x})$ are convex, then so is $f(\mathbf{x}) = \max_{\ell=1,\dots,L} f_\ell(\mathbf{x})$.
- Certain compositions of the form $f(\mathbf{x}) = g(h(\mathbf{x}))$ are convex under certain conditions on g and h (see Section 3.2 of Boyd and Vandenberghe's book)
  - Simplest case: If h is a linear (or affine) function and g is convex, then f is convex.

Jensen's inequality.

- *Jensen's inequality* states that, for any random vector $\mathbf{X}$ and convex function f, it holds that

$$f(\mathbb{E}[\mathbf{X}]) \leq \mathbb{E}[f(\mathbf{X})].$$

  This is used in countless proofs in machine learning, statistics, information theory, etc.

- Note that the inequality is true directly from the definition of convexity when $\mathbf{X}$ equals one value $\mathbf{x}$ with probability λ, and another value $\mathbf{x}'$ with probability $1 - \lambda$. Jensen's inequality states the more general form for an arbitrary distribution on $\mathbf{X}$.

2 Convex Optimization

- In machine learning and other fields, we are frequently interested in minimizing some cost function (or maximizing some utility function), possibly subject to certain constraints.

- We have already seen both constrained and unconstrained examples; recall the unconstrained form of the SVM:

$$\text{minimize}_{\boldsymbol{\theta},\theta_0} \quad \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{n}\left[1 - y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0)\right]_+, \tag{1}$$

where $[z]_+ = \max\{0, z\}$, and the constrained form of the SVM:

$$\text{minimize}_{\boldsymbol{\theta},\theta_0,\boldsymbol{\zeta}} \quad \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{n}\zeta_t \ \text{ subject to } \ y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) \geq 1 - \zeta_t \ \text{ and } \ \zeta_t \geq 0, \quad \forall t. \tag{2}$$

- We will return to the SVM later, but for now let's look at a more general optimization problem:

$$\begin{aligned} \text{minimize}_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m_{\text{ineq}} \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \dots, m_{\text{eq}}. \end{aligned} \tag{3}$$

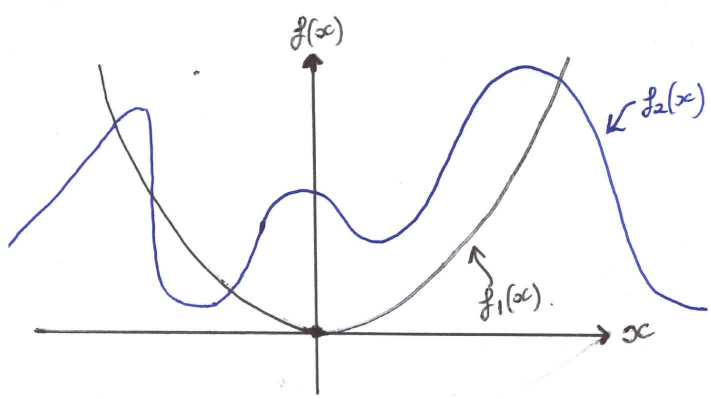
There are m_{ineq} *inequality constraints* and m_{eq} *equality constraints*.

  - Example: In (2) we have $\mathbf{x} = (\boldsymbol{\theta}, \theta_0, \boldsymbol{\zeta})$, $m_{\text{ineq}} = 2n$, and $m_{\text{eq}} = 0$, with the corresponding inequality constraint functions $f_i(\mathbf{x})$ being $1 - \zeta_t - y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0)$ and $-\zeta_t$ for $t = 1, \dots, n$.

- **Definition.** We say that (3) is a *convex optimization problem* if (i) $f_0(\mathbf{x})$ is convex; (ii) $f_i(\mathbf{x})$ is convex for all $i = 1, \dots, m_{\text{ineq}}$; (iii) $h_i(\mathbf{x})$ is affine for all $i = 1, \dots, m_{\text{eq}}$.

- This definition is very useful because, although solving (constrained or unconstrained) optimization problems is extremely hard in general, convexity is usually enough to permit finding a solution (sometimes analytically, but more often numerically).

- We can get some intuition by looking at the 1D case – which of these functions is easier to optimize using gradient descent techniques?



3 Lagrange Multipliers and Duality

- For an optimization problem of the form (3), the *Lagrangian* is defined as

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{m_{\text{eq}}} \nu_i h_i(\mathbf{x}), \tag{4}$$

where we have introduced extra parameters $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{m_{\text{ineq}}})$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_{m_{\text{eq}}})$. These are known as *Lagrange multipliers.*

  - We assume that $\lambda_i \geq 0$ for all i, whereas $\nu_i \in \mathbb{R}$ may be positive or negative.
  - Intuition: We no longer insist that $f_i(\mathbf{x}) \leq 0$, but we pay a penalty (scaled by λ_i) if it fails to hold. Conversely, we are "rewarded" if $f_i(\mathbf{x}) < 0$, i.e., strict inequality.

- **Important observation.** For any $\mathbf{x}$ feasible in (3), and any $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ with $\lambda_i \geq 0$, we have

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\mathbf{x}). \tag{5}$$

  - Proof: Follows immediately from $\lambda_i \geq 0$, $f_i(\mathbf{x}) \leq 0$, and $h_i(\mathbf{x}) = 0$.

- Minimizing both sides of (5) over $\mathbf{x}$ gives

$$\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\mathbf{x}^*), \tag{6}$$

where $\mathbf{x}^*$ is an optimal solution to (3).

  - The function

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

    is called the *Lagrange dual function.*

- Since $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ lower bounds $f_0(\mathbf{x}^*)$ according to (6), it is natural to look for the *best (highest) lower bound.* This leads to the *Lagrange dual problem*:

$$\begin{aligned} \text{maximize}_{\boldsymbol{\lambda}, \boldsymbol{\nu}} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} \quad & \lambda_i \geq 0, \qquad i = 1, \dots, m_{\text{ineq}}. \end{aligned} \tag{7}$$

Henceforth, let $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ denote the maximizer.

- **Duality.**
  - Since (6) holds for all $(\boldsymbol{\lambda}, \boldsymbol{\nu})$, it holds in particular for $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$, yielding

$$g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \leq f_0(\mathbf{x}^*).$$

    This is known as *weak duality.*
  - One of the most important results in convex optimization is that, if the original optimization problem is convex (i.e., f_0 and f_i are convex functions, and h_i is are linear functions), and a mild

regularity condition holds, then

$$g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = f_0(\mathbf{x}^*). \tag{8}$$

This is known as *strong duality*.

  * There are many possible "mild regularity conditions"; the most well-known is known as *Slater's condition*: There exists at least one point $\mathbf{x}$ in the relative interior of the domain satisfying the constraints of (3) with strict inequality (i.e., $f_i(\mathbf{x}) < 0$ and $h_i(\mathbf{x}) = 0$).
  * Another (more restrictive) sufficient condition is that the constraint functions f_i ($i = 1, \ldots, m_{\text{ineq}}$) are not only convex, but linear.

- Minimax theorem viewpoint: One way to understand duality is to interpret the original constrained optimization problem as solving

$$\min_{\mathbf{x}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

This is because the inner maximization (more precisely a supremum) equals ∞ whenever $f_i(\mathbf{x}) > 0$ or $h_i(\mathbf{x}) \neq 0$, because any arbitrarily large value can be achieved by taking the corresponding λ_i or ν_i to be huge. In addition, when $\mathbf{x}$ satisfies the constraints (i.e., each $f_i(\mathbf{x}) \leq 0$ and $h_i(\mathbf{x}) = 0$), it is not hard to show that that $\max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x})$ (achieved by $\boldsymbol{\lambda} = \mathbf{0}$ and $\boldsymbol{\nu} = \mathbf{0}$).

In contrast, the Lagrange dual problem solves

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

So it's the same problem, just with the max and min swapped!

It a well-known fact of optimization and game theory that $\min_A \max_B f(A, B) \geq \max_B \min_A f(A, B)$. Strong duality is related to the *minimax theorem* (see `https://en.wikipedia.org/wiki/Minimax_theorem`), which states that in fact

$$\min_A \max_B f(A, B) = \max_B \min_A f(A, B)$$

in the case that $f(A, \cdot)$ is concave in B and $f(\cdot, B)$ is convex in A.

- **Example (Linear programming).**

  - Consider a linear program of the form

$$\text{maximize}_{\mathbf{x}} \qquad \mathbf{c}^T \mathbf{x} \tag{9}$$

$$\text{subject to} \qquad \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} \tag{10}$$

for some matrix $\mathbf{A} \in \mathbb{R}^{m \times d}$ and vectors $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{c} \in \mathbb{R}^d$. The inequality $\mathbf{x} \geq \mathbf{0}$ should be interpreted as holding element-wise.

- Interpreting this as being in the form (3) with $m_{\text{ineq}} = m$ and $m_{\text{eq}} = d$, we have the Lagrangian

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) &= -\mathbf{c}^T\mathbf{x} - \sum_{i=1}^{d} \lambda_i x_i + \sum_{i=1}^{m} \nu_i(\mathbf{a}_i^T\mathbf{x} - b_i) \\ &= -\mathbf{c}^T\mathbf{x} - \boldsymbol{\lambda}^T\mathbf{x} + \boldsymbol{\nu}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= -\mathbf{b}^T\boldsymbol{\nu} + (\mathbf{A}^T\boldsymbol{\nu} - \boldsymbol{\lambda} - \mathbf{c})^T\mathbf{x}, \end{aligned}$$

  where $\mathbf{a}_i$ is the i-th row of $\mathbf{A}$, b_i is the i-th entry of $\mathbf{b}$, etc.

  * Note: Switching from "maximize" to "minimize" requires taking $f_0(\mathbf{x}) = -\mathbf{c}^T\mathbf{x}$

- Minimizing over $\mathbf{x}$, we find that $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ (which we recall is $\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$) takes the form

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \begin{cases} -\mathbf{b}^T\boldsymbol{\nu} & \mathbf{A}^T\boldsymbol{\nu} - \boldsymbol{\lambda} - \mathbf{c} = \mathbf{0} \\ -\infty & \text{otherwise.} \end{cases}$$

  This is because whenever $\mathbf{A}^T\boldsymbol{\nu} + \boldsymbol{\lambda} + \mathbf{c} \neq 0$, one can just make a suitable entry of x_i arbitrarily large in either the positive or negative direction.

- Substituting this expression for $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ into (7) yields the *dual problem*:

$$\begin{aligned} \text{maximize}_{\boldsymbol{\lambda}, \boldsymbol{\nu}} \quad & -\mathbf{b}^T\boldsymbol{\nu} \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mathbf{A}^T\boldsymbol{\nu} - \boldsymbol{\lambda} - \mathbf{c} = \mathbf{0}, \end{aligned}$$

  where the second constraint can be introduced since all other values yield a (certainly suboptimal) value of $-\infty$. Since $\boldsymbol{\lambda}$ does not appear in the objective function, we can further simplify the above maximization to

$$\begin{aligned} \text{minimize}_{\boldsymbol{\nu}} \quad & \mathbf{b}^T\boldsymbol{\nu} \\ \text{subject to} \quad & \mathbf{A}^T\boldsymbol{\nu} \geq \mathbf{c}. \end{aligned}$$

- If we replace $\mathbf{A}\mathbf{x} = \mathbf{b}$ by $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ in the original formulation, then we arrive at a similar dual expression but with the added constraint $\boldsymbol{\nu} \geq 0$.

- **An intuitive interpretation:**

  * The original problem constrains $\mathbf{A}\mathbf{x} = \mathbf{b}$; multiplying both sides on the left by $\boldsymbol{\nu}^T$ gives $\boldsymbol{\nu}^T\mathbf{A}\mathbf{x} = \boldsymbol{\nu}^T\mathbf{b}$, or equivalently $(\mathbf{A}^T\boldsymbol{\nu})^T\mathbf{x} = \mathbf{b}^T\boldsymbol{\nu}$ (by standard properties of the transpose)
  * Now, since $\mathbf{x} \geq \mathbf{0}$ and we are maximizing $\mathbf{c}^T\mathbf{x}$, we find that if $\mathbf{A}^T\boldsymbol{\nu} \geq \mathbf{c}$, it holds that $(\mathbf{A}^T\boldsymbol{\nu})^T\mathbf{x}$ is at least as high as $\mathbf{c}^T\mathbf{x}$. Then, by the previous dot point, $\mathbf{b}^T\boldsymbol{\nu}$ is at least as high as $\mathbf{c}^T\mathbf{x}$.
  * Hence, for any $\boldsymbol{\nu}$ that satisfies $\mathbf{A}^T\boldsymbol{\nu} \geq \mathbf{c}$, we have that $\mathbf{b}^T\boldsymbol{\nu}$ is at least as high as the original problem's optimal value, i.e., it is an *upper bound* to the optimal value.
  * By minimizing over all such $\boldsymbol{\nu}$ (as is done in the dual expression), we are finding the *lowest (best) possible upper bound*, and this turns out to make the upper bound hold with equality.

4 The Karush-Kuhn-Tucker (KKT) Conditions

- In the case that strong duality holds as per (8), we have the following chain of inequalities:

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\ &= \min_{\mathbf{x}} \left\{ f_0(\mathbf{x}) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^{m_{\text{eq}}} \nu_i^* h_i(\mathbf{x}) \right\} \\ &\le f_0(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{eq}}} \nu_i^* h_i(\mathbf{x}^*) \\ &\le f_0(\mathbf{x}^*), \end{aligned}$$

  where we first applied the definition of g, then upper bounded the minimum by the specific value $\mathbf{x}^*$, then used the fact that $f_i(\mathbf{x}^*) \le 0$ and $h_i(\mathbf{x}^*) = 0$.

- Since we ended up with $f_0(\mathbf{x}^*) \le f_0(\mathbf{x}^*)$, both of the inequalities must hold with equality. Let's look at these in more detail:

  – The first inequality holding with equality gives

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f_0(\mathbf{x}) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^{m_{\text{eq}}} \nu_i^* h_i(\mathbf{x}).$$

    Assuming the functions are differentiable, the fact that $\mathbf{x}^*$ is a minimizer means that the derivative must vanish:

$$\nabla f_0(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{eq}}} \nu_i^* \nabla h_i(\mathbf{x}^*) = \mathbf{0}.$$

  – The second inequality holding with equality gives

$$\lambda_i^* f_i(\mathbf{x}^*) = 0, \qquad i = 1, \dots, m_{\text{ineq}}.$$

    This means that either $f_i(\mathbf{x}^*) = 0$ (i.e., the constraint holds with equality) or $\lambda_i^* = 0$. This property is known as *complementary slackness*.

- Summarizing the above leads to a set of conditions on $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ known as the *KKT conditions*:

  1. (Primal feasibility) $f_i(\mathbf{x}^*) \le 0$ for $i = 1, \dots, m_{\text{ineq}}$, and $h_i(\mathbf{x}^*) = 0$ for $i = 1, \dots, m_{\text{eq}}$.
  2. (Dual feasibility) $\lambda_i^* \ge 0$ for $i = 1, \dots, m_{\text{ineq}}$.
  3. (Complementary slackness) $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for $i = 1, \dots, m_{\text{ineq}}$.
  4. (Vanishing gradient) $\nabla f_0(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{eq}}} \nu_i^* \nabla h_i(\mathbf{x}^*) = \mathbf{0}$.

  These generalize the requirement that the *unconstrained* maximizer of $f_0(\mathbf{x})$ should satisfy $\nabla f_0(\mathbf{x}^*) = 0$.

  – General case: If strong duality holds, it is necessary that $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ satisfy the KKT conditions.

  – Convex case: If strong duality holds *and the primal problem is convex*, then $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ satisfying the KKT conditions are *also sufficient for optimality* (the proof of this is omitted).

5 Support Vector Machine Revisited

Forming the dual expression.

- We have seen a few equivalent "primal" formulations of SVM; let's consider the following one (focusing on the hard-margin formulation with offset for now):

$$\text{minimize}_{\boldsymbol{\theta},\theta_0} \ \frac{1}{2}\|\boldsymbol{\theta}\|^2 \ \text{ subject to } \ y_t(\boldsymbol{\theta}^T\mathbf{x}_t+\theta_0)\geq 1, \ \ \forall t=1,\dots,n. \tag{11}$$

  This is a convex optimization problem with affine constraints, so strong duality holds.

- The Lagrangian is given by

$$L(\boldsymbol{\theta},\theta_0,\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + \sum_{t=1}^n \lambda_t\big(1-y_t(\boldsymbol{\theta}^T\mathbf{x}_t+\theta_0)\big). \tag{12}$$

- To find $g(\boldsymbol{\lambda}) = \min_{\mathbf{x}} L(\boldsymbol{\theta},\theta_0,\boldsymbol{\lambda})$, we set the partial derivatives to zero. For θ_0:

$$\frac{\partial L}{\partial \theta_0} = -\sum_{t=1}^n \lambda_t y_t = 0, \tag{13}$$

  and for $\boldsymbol{\theta}$ (with a bit of basic vector calculus):

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \boldsymbol{\theta} - \sum_{t=1}^n \lambda_t y_t \mathbf{x}_t = \mathbf{0},$$

  which implies $\boldsymbol{\theta} = \boldsymbol{\theta}^* := \sum_{t=1}^n \lambda_t y_t \mathbf{x}_t$.

- Under these optimality conditions, the second term in (12) simplifies to

$$\sum_{t=1}^n \lambda_t\big(1-y_t(\boldsymbol{\theta}^T\mathbf{x}_t+\theta_0)\big) = \sum_{t=1}^n \lambda_t - \sum_{t=1}^n \lambda_t y_t \bigg(\sum_{s=1}^n \lambda_s y_s \mathbf{x}_s\bigg)^T \mathbf{x}_t \tag{14}$$

$$= \sum_{t=1}^n \lambda_t - \bigg(\sum_{s=1}^n \lambda_s y_s \mathbf{x}_s\bigg)^T \bigg(\sum_{t=1}^n \lambda_t y_t \mathbf{x}_t\bigg). \tag{15}$$

  But $\|\boldsymbol{\theta}\|^2$ with $\boldsymbol{\theta} = \sum_{t=1}^n \lambda_t y_t \mathbf{x}_t$ is also equal to $\big(\sum_{s=1}^n \lambda_s y_s \mathbf{x}_s\big)^T\big(\sum_{t=1}^n \lambda_t y_t \mathbf{x}_t\big)$, so overall (12) gives

$$g(\boldsymbol{\lambda}) = L(\boldsymbol{\theta}^*,\theta_0^*,\boldsymbol{\lambda}) = \begin{cases} \sum_{t=1}^n \lambda_t - \frac{1}{2}\sum_{s=1}^n\sum_{t=1}^n \lambda_s\lambda_t y_s y_t \mathbf{x}_s^T\mathbf{x}_t & \sum_{t=1}^n \lambda_t y_t = 0 \\ -\infty & \text{otherwise.} \end{cases} \tag{16}$$

  The first case can be thought of as corresponding to (13), and the second case results because if $\sum_{t=1}^n \lambda_t y_t \neq 0$ then one can choose θ_0 arbitrarily large (in the positive or negative direction) to make the right-hand side of (12) arbitrarily negative.

- Renaming $\boldsymbol{\lambda}$ as $\boldsymbol{\alpha}$ and maximizing the Lagrange dual function g, we arrive at the *dual formulation of* *the SVM (separable case)*:

$$\begin{aligned}\text{maximize}_{\boldsymbol{\alpha}} \quad & \sum_{t=1}^{n} \alpha_t - \frac{1}{2}\sum_{s=1}^{n}\sum_{t=1}^{n} \alpha_s \alpha_t y_s y_t \mathbf{x}_s^T \mathbf{x}_t \\ \text{subject to} \quad & \alpha_t \geq 0 \;\; \forall t \in \{1,\dots,n\}, \\ & \sum_{t=1}^{n} \alpha_t y_t = 0.\end{aligned}$$

Observe that $-\infty$ case with $\sum_{t=1}^n \alpha_t y_t = 0$ was turned into a constraint, on the basis that a value of $-\infty$ can never be optimal.

Recovering the classifier.

- We have already shown that $\boldsymbol{\theta} = \sum_{t=1}^n \alpha_t y_t \mathbf{x}_t$, so to form a classifier we only need to find θ_0.
- By the complementary slackness condition in the KKT conditions, each training sample falls into one of the following categories:[7]
  - (Support vectors) $\alpha_t > 0$ and $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) = 1$, i.e., the point is on the margin's boundary;
  - (Not support vectors) $\alpha_t = 0$ and $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) > 1$, i.e., the point is outside the margin.

  To find θ_0, we can just take any $(\mathbf{x}_t, y_t)$ corresponding to the former case, and set $\theta_0 = \frac{1}{y_t} - \boldsymbol{\theta}^T\mathbf{x}_t$.

Interpretation of the support vector property.

- In the SVM lecture, we stated that the maximum margin is determined only by the support vectors, and re-running SVM with the non-support-vectors removed leads to exactly the same decision boundary and margins.
- This can be understood better via the theory of convex analysis by observing that the support vectors are exactly those with Lagrange multiplier $\alpha_t > 0$ (as stated above).
- In convex analysis, a Lagrange multiplier of zero corresponds to an *inactive constrained* – one which, when removed, does not change the optimal solution (e.g., consider the problem "minimize z^2 subject to $z \geq -1$", whose solution is attained with $z = 0$).
- Removing a non-support-vector from the data set amounts to removing its constraint from the (primal) SVM optimization formulation. But since the Lagrange multiplier is zero, this does not change the optimal solution.
- (This discussion is not a formal proof, but highlights that this support vector property is a special case of a general phenomenon in convex analysis)

Soft-margin formulation and kernel SVM.

- For the soft-margin SVM, a similar analysis via Lagrange duality reveals that the primal formulation

$$\text{minimize}_{\boldsymbol{\theta},\theta_0,\boldsymbol{\zeta}} \;\; \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{n}\zeta_t \;\; \text{subject to} \;\; y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) \geq 1 - \zeta_t \;\; \text{and} \;\; \zeta_t \geq 0, \quad \forall t \tag{17}$$

[7]In principle, complementary slackness also allows both $\alpha_t = 0$ and $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) = 1$ to hold simultaneously, but we will not worry about this unusual case.

has a dual formulation given by

$$\begin{aligned}
\text{maximize}_{\boldsymbol{\alpha}} \qquad & \sum_{t=1}^n \alpha_t - \frac{1}{2}\sum_{s=1}^n\sum_{t=1}^n \alpha_s\alpha_t y_s y_t \mathbf{x}_s^T\mathbf{x}_t \\
\text{subject to} \qquad & \alpha_t \in [0, C] \ \ \forall t \in \{1,\dots,n\}, \\
& \sum_{t=1}^n \alpha_t y_t = 0.
\end{aligned}$$

This is exactly the same as above, except for the added constraint $\alpha_t \le C$. (And as we have seen before, taking $C \to \infty$ recovers the hard-margin formulation)

- The dual variables are only slightly trickier to understand in this case; one can use complementary slackness to show the following:[8]
  - If some $\mathbf{x}_t$ is "strictly" on the correct side of the margin (i.e., $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) > 1$), then $\alpha_t = 0$.
  - If some $\mathbf{x}_t$ is inside the margin or mis-classified (i.e., $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) < 1$), then $\alpha_t = C$.
  - If $0 < \alpha_t < C$, then $\mathbf{x}_t$ is exactly on the margin (i.e., $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) = 1$).

  Hence, $\mathbf{x}_t$ is a support vector if and only if $\alpha_t > 0$, just like in the separable case.

- The final classifier applied to $\mathbf{x}$ is given by

$$\text{sign}(\boldsymbol{\theta}^T\mathbf{x} + \theta_0) = \text{sign}\bigg(\sum_{t=1}^n \alpha_t y_t \mathbf{x}_t^T\mathbf{x} + \theta_0\bigg),$$

  where we have again applied $\boldsymbol{\theta} = \sum_{t=1}^n \alpha_t y_t \mathbf{x}_t$ (previously written as $\boldsymbol{\theta} = \sum_{t=1}^n \lambda_t y_t \mathbf{x}_t$).
  - To find θ_0, we take any $(\mathbf{x}_t, y_t)$ corresponding to a t with $0 < \alpha_t < C$, and set $\theta_0 = \frac{1}{y_t} - \boldsymbol{\theta}^T\mathbf{x}_t$.
  - Because strong duality holds, the resulting classifier is *identical* to the primal SVM classifier

- In both the separable and non-separable case, the classifier depends on $\{\mathbf{x}_t\}_{t=1}^n$ only through the inner products $\langle \mathbf{x}_s, \mathbf{x}_t\rangle = \mathbf{x}_s^T\mathbf{x}_t$, so we can apply the *kernel trick*, leading to the following:

- Kernel SVM:[9] Find $\boldsymbol{\alpha}$ by solving the optimization problem

$$\begin{aligned}
\text{maximize}_{\boldsymbol{\alpha}} \qquad & \sum_{t=1}^n \alpha_t - \frac{1}{2}\sum_{s=1}^n\sum_{t=1}^n \alpha_s\alpha_t y_s y_t k(\mathbf{x}_s, \mathbf{x}_t) \\
\text{subject to} \qquad & \alpha_t \in [0, C] \ \ \forall t \in \{1,\dots,n\}, \\
& \sum_{t=1}^n \alpha_t y_t = 0.
\end{aligned}$$

  and let the final classification rule be

$$\hat{y}(\mathbf{x}) = \text{sign}\bigg(\sum_{t=1}^n \alpha_t y_t k(\mathbf{x}, \mathbf{x}_t) + \theta_0\bigg),$$

[8] See the tutorial for the relevant analysis. Again, in principle when $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) = 1$ any $\alpha_t \in [0, C]$ could still be allowed, and not just $\alpha_t \in (0, C)$. On the other hand, $\alpha_t \in (0, C)$ definitively implies that $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) = 1$.

[9] This is based on the dual SVM formulation. The primal formulation doesn't lend itself directly to the kernel trick, but it is possible to obtain a primal-type kernel SVM formulation (without needing Lagrange duality) using a result called the *Representer Theorem*. See Section 16.2 of the "Understanding Machine Learning" book for details.

where θ_0 is found in the same way as above, replacing $\theta_0 = \frac{1}{y_t} - \boldsymbol{\theta}^T \mathbf{x}_t$ by $\theta_0 = \frac{1}{y_t} - \sum_{s=1}^n \alpha_s y_s k(\mathbf{x}_s, \mathbf{x}_t)$.

Computational considerations.

- The choice of whether to use the primal or dual formulation is often dictated by computational considerations, particularly for large d ("high-dimensional") and/or large n ("big data")
- The bottleneck in the dual formulation is often computing $O(n^2)$ values of $k(\mathbf{x}_s, \mathbf{x}_t)$