HT2015: SC4
Statistical Data Mining and Machine Learning

Dino Sejdinovic
Department of Statistics
Oxford

http://www.stats.ox.ac.uk/~sejdinov/sdmml.html

Convex Optimization and Support Vector Machines

slides based on Arthur Gretton's Advanced Topics in Machine Learning course

Optimization and the Lagrangian

Optimization problem on $x \in \mathbb{R}^d$ / primal,

$$\begin{aligned} \text{minimize} \quad & f_0(x) & \\ \text{subject to} \quad & f_i(x) \leq 0 & i = 1, \ldots, m \\ & h_j(x) = 0 & j = 1, \ldots r. \end{aligned}$$

- domain $\mathcal{D} := \bigcap_{i=0}^{m} \mathrm{dom} f_i \; \cap \; \bigcap_{j=1}^{r} \mathrm{dom} h_j$ (nonempty).
- p^*: the (primal) optimal value

Idealy we would want an unconstrained problem

$$\text{minimize}\, f_0(x) + \sum_{i=1}^{m} I_-\left(f_i(x)\right) + \sum_{j=1}^{r} I_0\left(h_j(x)\right),$$

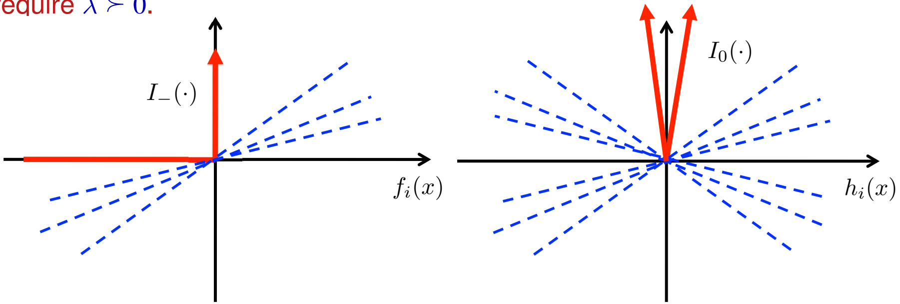
where $I_-(u) = \begin{cases} 0, & u \leq 0 \\ \infty, & u > 0 \end{cases}$ and $I_0(u) = \begin{cases} 0, & u = 0 \\ \infty, & u \neq 0 \end{cases}$.

Lower bound interpretation of Lagrangian

The **Lagrangian** $L : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ is an (easier to optimize) lower bound on the original problem:

$$L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^{m} \underbrace{\lambda_i f_i(x)}_{\leq I_-(f_i(x))} + \sum_{j=1}^{r} \underbrace{\nu_j h_j(x)}_{\leq I_0(h_j(x))},$$

and has domain $\mathrm{dom} L := \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^r$. The vectors λ and ν are called **Lagrange multipliers** or **dual variables**. To ensure a lower bound, we require $\lambda \succeq 0$.



Lagrange dual: lower bound on optimum p^*

The **Lagrange dual function:** minimize Lagrangian When $\lambda \succeq 0$ and $f_i(x) \leq 0$, Lagrange dual function is

$$g(\lambda, \nu) := \inf_{x \in \mathcal{D}} L(x, \lambda, \nu).$$

A **dual feasible** pair (λ, ν) is a pair for which $\lambda \succeq 0$ and $(\lambda, \nu) \in \mathrm{dom}(g)$.
We will show: (next slide) for any $\lambda \succeq 0$ and ν,

$$g(\lambda, \nu) \leq f_0(x)$$

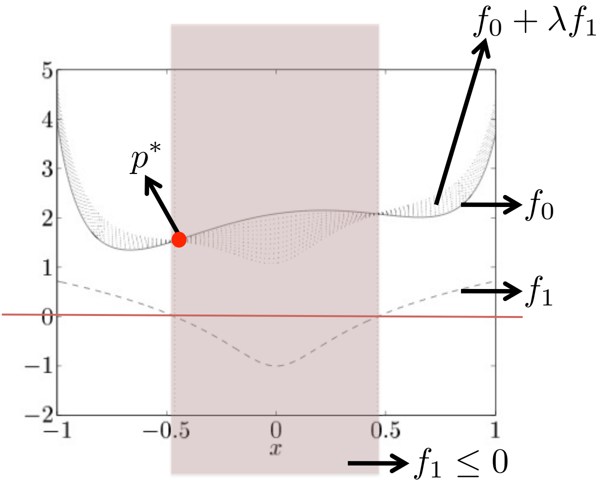
wherever

$$\begin{aligned} f_i(x) &\leq 0 \\ h_j(x) &= 0 \end{aligned}$$

(including at optimal point $f_0(x^*) = p^*$).

Lagrange dual: lower bound on optimum p^*

Simplest example: minimize over x the function $L(x, \lambda) = f_0(x) + \lambda f_1(x)$



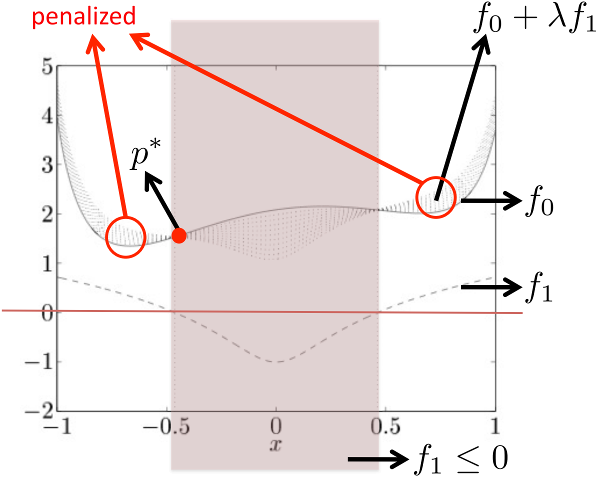
Reminders:

- f_0 is function to be minimized.
- $f_1 \leq 0$ is inequality constraint
- $\lambda \geq 0$ is Lagrange multiplier
- p^* is minimum f_0 **in constraint set**

Figure from Boyd and Vandenberghe

Lagrange dual: lower bound on optimum p^*

Simplest example: minimize over x the function $L(x, \lambda) = f_0(x) + \lambda f_1(x)$



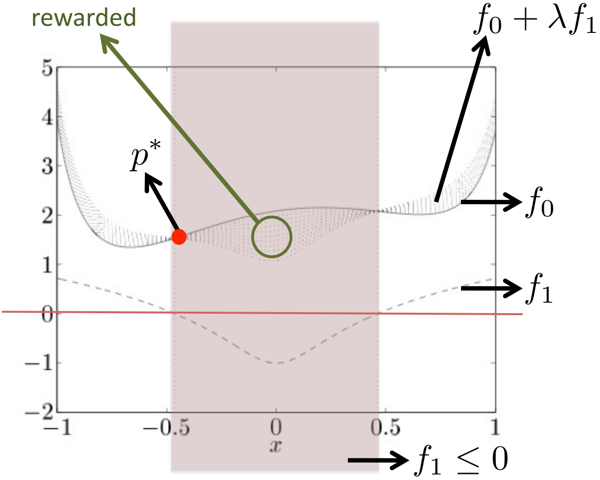
Reminders:

- f_0 is function to be minimized.
- $f_1 \leq 0$ is inequality constraint
- $\lambda \geq 0$ is Lagrange multiplier
- p^* is minimum f_0 **in constraint set**

Figure from Boyd and Vandenberghe

Lagrange dual: lower bound on optimum p^*

Simplest example: minimize over x the function $L(x, \lambda) = f_0(x) + \lambda f_1(x)$



Reminders:

- f_0 is function to be minimized.
- $f_1 \leq 0$ is inequality constraint
- $\lambda \geq 0$ is Lagrange multiplier
- p^* is minimum f_0 **in constraint set**

Figure from Boyd and Vandenberghe

Lagrange dual is a lower bound on p^*

Assume $\tilde{x}$ is feasible, i.e. $f_i(\tilde{x}) \leq 0$, $h_i(\tilde{x}) = 0$, $\tilde{x} \in \mathcal{D}$, $\lambda \succeq 0$. Then

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{r} \nu_i h_i(\tilde{x}) \leq 0$$

Thus

$$\begin{aligned} g(\lambda, \nu) \quad &:= \quad \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{r} \nu_i h_i(x) \right) \\ &\leq \quad f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{r} \nu_i h_i(\tilde{x}) \\ &\leq \quad f_0(\tilde{x}). \end{aligned}$$

This holds for every feasible $\tilde{x}$, hence lower bound holds.

Best lower bound: maximize the dual

Best lower bound $g(\lambda, \nu)$ on the optimal solution p^* of original problem:
Lagrange dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0. \end{array}$$

Dual feasible: (λ, ν) with $\lambda \succeq 0$ and $g(\lambda, \nu) > -\infty$.
Dual optimal: solutions (λ^*, ν^*) to the dual problem, d^* is optimal value (**dual always easy to maximize**: next slide).
Weak duality always holds:

$$\max_{\lambda \succeq 0, \nu} \underbrace{\min_{x \in \mathcal{D}} L(x, \lambda, \nu)}_{= g(\lambda, \nu)} = d^* \leq p^* = \min_{x \in \mathcal{D}} \underbrace{\max_{\lambda \succeq 0, \nu} L(x, \lambda, \nu)}_{= \begin{cases} f_0(x) & \text{if constraints satisfied,} \\ \infty & \text{otherwise.} \end{cases}}$$

Strong duality: (does **not** always hold, conditions given later):

$$d^* = p^*.$$

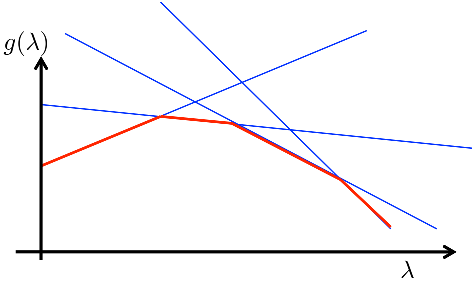
If strong duality holds: solve the **easy (concave) dual problem** to find p^*.

Maximizing the dual is always easy

The **Lagrange dual function:** minimize Lagrangian (lower bound)

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu).$$

Dual function is a pointwise infimum of affine functions of (λ, ν), hence **concave** in (λ, ν) with convex constraint set $\lambda \succeq 0$.



Example:

One inequality constraint,

$$L(x, \lambda) = f_0(x) + \lambda f_1(x),$$

and assume there are only four possible values for x. Each line represents a different x.

How do we know if strong duality holds?

Conditions under which strong duality holds are called **constraint qualifications** (they are sufficient, but not necessary)

(Probably) best known sufficient condition: Strong duality holds if

- Primal problem is **convex**, i.e. of the form

$$
\begin{aligned}
\text{minimize} \quad & f_0(x) & \\
\text{subject to} \quad & f_i(x) \leq 0 & i = 1, \ldots, n \\
& Ax = b &
\end{aligned}
$$

  for convex $f_0, \ldots, f_m$, **and**

Slater's condition: there exists a strictly feasible point $\tilde{x}$, such that $f_i(\tilde{x}) < 0$, $i = 1, \ldots, n$ (reduces to the existence of any feasible point when inequality constraints are affine, i.e., $Cx \preceq d$).

A consequence of strong duality...

Assume primal is equal to the dual. What are the consequences?

- x^* solution of original problem (minimum of f_0 under constraints),
- (λ^*, ν^*) solutions to dual

$$
\begin{aligned}
f_0(x^*) &\underset{\text{(assumed)}}{=} g(\lambda^*, \nu^*) \\
&\underset{\text{(g definition)}}{=} \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \right) \\
&\underset{\text{(inf definition)}}{\leq} f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \\
&\underset{(4)}{\leq} f_0(x^*),
\end{aligned}
$$

(4): (x^*, λ^*, ν^*) satisfies $\lambda^* \succeq 0$, $f_i(x^*) \leq 0$, and $h_i(x^*) = 0$.

...is complementary slackness

From previous slide,

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0, \tag{1}$$

which is the condition of **complementary slackness**. This means

$$\begin{aligned} \lambda_i^* > 0 \quad &\Longrightarrow \quad f_i(x^*) = 0, \\ f_i(x^*) < 0 \quad &\Longrightarrow \quad \lambda_i^* = 0. \end{aligned}$$

From λ_i, read off which inequality constraints are strict.

KKT conditions for global optimum

Assume functions f_i, h_i are differentiable and **strong duality**. Since x^* minimizes $L(x, \lambda^*, \nu^*)$, derivative at x^* is zero,

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{r} \nu_i^* \nabla h_i(x^*) = 0.$$

KKT conditions definition: we are at **global optimum,** (x^*, λ^*, ν^*) when (a) **strong duality** holds, and (b):

$$\begin{aligned}
f_i(x^*) &\leq 0,\ i = 1, \ldots, m \\
h_i(x^*) &= 0,\ i = 1, \ldots, r \\
\lambda_i^* &\geq 0,\ i = 1, \ldots, m \\
\lambda_i^* f_i(x^*) &= 0,\ i = 1, \ldots, m \\
\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{r} \nu_i^* \nabla h_i(x^*) &= 0
\end{aligned}$$

KKT conditions for global optimum

In summary: if

- primal problem convex and
- inequality constraints affine

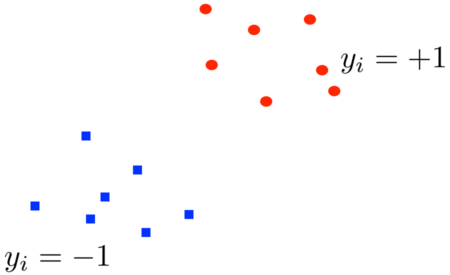
then strong duality holds. If in addition

- functions f_i, h_i differentiable

then KKT conditions are **necessary and sufficient** for optimality.

Linearly separable points

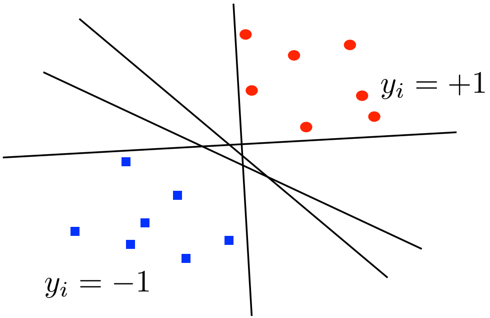
Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.



Data given by $\{x_i, y_i\}_{i=1}^{n}$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1, +1\}$

Linearly separable points

Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.

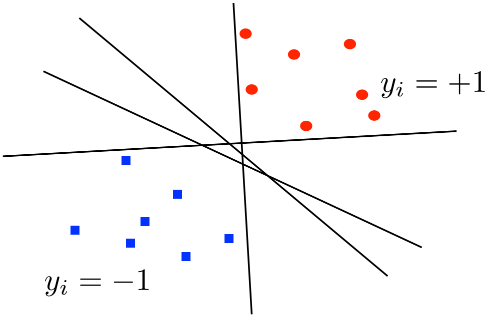


Hyperplane equation $w^\top x + b = 0$. Linear discriminant given by

$$f(x) = \operatorname{sign}(w^\top x + b)$$

Linearly separable points

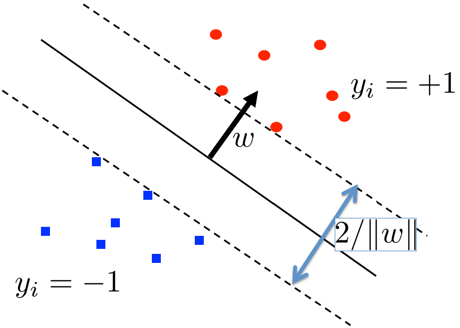
Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.



For a datapoint close to the decision boundary, a small change leads to a change in classification. Can we make the classifier more robust?

Linearly separable points

Classify two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.



Smallest distance from each class to the separating hyperplane $w^\top x + b$ is called the **margin**.

Maximum margin classifier, linearly separable case

This problem can be expressed as follows:

$$\max_{w,b}\,(\text{margin}) = \max_{w,b}\left(\frac{1}{\|w\|}\right)$$

subject to

$$\begin{cases} w^\top x_i + b \geq 1 & i\ :\ y_i = +1, \\ w^\top x_i + b \leq -1 & i\ :\ y_i = -1. \end{cases}$$

The resulting classifier is

$$f(x) = \operatorname{sign}(w^\top x + b),$$

We can rewrite to obtain a **quadratic program**:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

subject to

$$y_i(w^\top x_i + b) \geq 1.$$

Maximum margin classifier: with errors allowed

Allow “errors”: points within the margin, or even on the wrong side of the decision boudary. Ideally:

$$\min_{w,b}\left(\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\mathbb{I}[y_i\left(w^\top x_i+b\right)<0]\right),$$

where C controls the tradeoff between maximum margin and loss.
Replace with **convex upper bound**:

$$\min_{w,b}\left(\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}h\left(y_i\left(w^\top x_i+b\right)\right)\right).$$

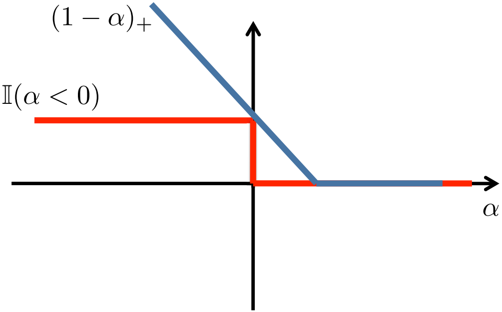
with hinge loss,

$$h(\alpha)=(1-\alpha)_+=\begin{cases}1-\alpha, & 1-\alpha>0\\ 0, & \text{otherwise.}\end{cases}$$

Hinge loss

Hinge loss:

$$h(\alpha) = (1-\alpha)_+ = \begin{cases} 1-\alpha, & 1-\alpha > 0 \\ 0, & \text{otherwise.} \end{cases}$$



Support vector classification

Substituting in the hinge loss, we get

$$\min_{w,b}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} h\left(y_i\left(w^\top x_i + b\right)\right)\right).$$

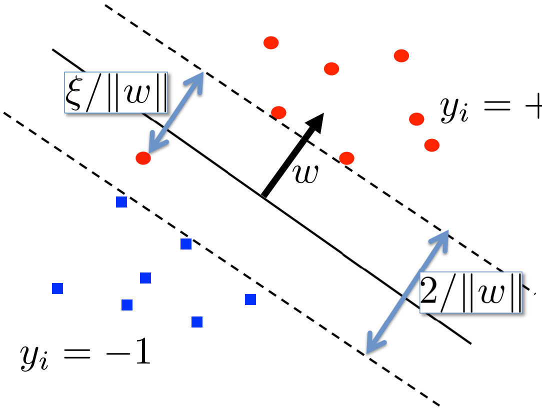
To simplify, use substitution $\xi_i = h\left(y_i\left(w^\top x_i + b\right)\right)$:

$$\min_{w,b,\xi}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right)$$

subject to

$$\xi_i \geq 0 \qquad y_i\left(w^\top x_i + b\right) \geq 1 - \xi_i$$

Support vector classification



Does strong duality hold?

1. Is the optimization problem convex wrt the variables w, b, ξ?

$$\begin{aligned} \text{minimize} \quad & f_0(w,b,\xi) := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & f_i(w,b,\xi) := 1 - \xi_i - y_i\left(w^\top x_i + b\right) \leq 0,\ i = 1,\ldots,n \\ & f_i(w,b,\xi) := -\xi_i \leq 0,\ i = n+1,\ldots,2n \end{aligned}$$

   Each of $f_0, f_1, \ldots, f_n$ are convex. No equality constraints.

2. Does Slater's condition hold? Yes (trivially) since inequality constraints **affine**.

Thus **strong duality** holds, the problem is differentiable, hence the KKT conditions hold at the global optimum.

Support vector classification: Lagrangian

The Lagrangian: $L(w, b, \xi, \alpha, \lambda) =$

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - y_i\left(w^\top x_i + b\right)\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i)$$

with dual variable constraints

$$\alpha_i \geq 0, \qquad \lambda_i \geq 0.$$

Minimize wrt the primal variables w, b, and ξ.
Derivative wrt w:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \qquad w = \sum_{i=1}^{n}\alpha_i y_i x_i.$$

Derivative wrt b:

$$\frac{\partial L}{\partial b} = \sum_i y_i \alpha_i = 0.$$

Support vector classification: Lagrangian

Derivative wrt ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0 \qquad \alpha_i = C - \lambda_i.$$

Since $\lambda_i \geq 0$,

$$\alpha_i \leq C.$$

Now use complementary slackness:

Non-margin SVs (margin errors): $\alpha_i = C > 0$:

1. We immediately have $y_i \left(w^\top x_i + b \right) = 1 - \xi_i$.
2. Also, from condition $\alpha_i = C - \lambda_i$, we have $\lambda_i = 0$, so $\xi_i \geq 0$

Margin SVs: $0 < \alpha_i < C$:

1. We again have $y_i \left(w^\top x_i + b \right) = 1 - \xi_i$.
2. This time, from $\alpha_i = C - \lambda_i$, we have $\lambda_i > 0$, hence $\xi_i = 0$.

Non-SVs (on the correct side of the margin): $\alpha_i = 0$:

1. From $\alpha_i = C - \lambda_i$, we have $\lambda_i > 0$, hence $\xi_i = 0$.
2. Thus, $y_i \left(w^\top x_i + b \right) \geq 1$

The support vectors

We observe:

1. The solution is sparse: points which are neither on the margin nor "margin errors" have $\alpha_i = 0$
2. The support vectors: only those points on the decision boundary, or which are margin errors, contribute.
3. Influence of the non-margin SVs is bounded, since their weight cannot exceed C.

Support vector classification: dual function

Thus, our goal is to maximize the dual,

$$
\begin{aligned}
g(\alpha, \lambda) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(w^\top x_i + b\right) - \xi_i\right) \\
&\quad + \sum_{i=1}^{n}\lambda_i(-\xi_i) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j \\
&\quad \underbrace{-b\sum_{i=1}^{n}\alpha_i y_i}_{0} + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}(C-\alpha_i)\xi_i \\
&= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j.
\end{aligned}
$$

Support vector classification: dual problem

Maximize the dual,

$$g(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j,$$

subject to the constraints

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} y_i \alpha_i = 0$$

This is a quadratic program. From α, obtain the hyperplane with $w = \sum_{i=1}^{n} \alpha_i y_i x_i$

Offset b can be obtained from any of the margin SVs: $1 = y_i \left(w^\top x_i + b \right)$.