IE 506 - Machine Learning: Principles and Techniques

SVMs and Duality

February 20, 2024.

SVM - primal problem formulation

Soft-margin Support Vector Machine

Optimization Problem Formulation for Soft-margin SVM

Given a data set $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^n$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{+1, -1\}$, soft-margin SVM solves:

$$\min_{w,b,\xi_i \geq 0} \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{s.t. } y^i(\langle w, x^i \rangle - b) \geq 1 - \xi_i, \ \forall i \in \{1, 2, \ldots, n\}. \tag{1}$$

- $C > 0$ is a regularization constant.
- $\xi_i \geq 0$ denotes the slack associated with the data point (x^i, y^i).
  - If x^i lies on the **improper** side of the separating hyperplane, the slack ξ_i is positive.
  - If x^i lies on the **proper** side of the separating hyperplane, the slack ξ_i is simply zero.

Soft-margin Support Vector Machine

Note that the soft-margin SVM can be equivalently written as:

$$\begin{aligned} \min_{w,b,\xi_i \geq 0} \ & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \ & \left\langle w, x^i \right\rangle \geq b + 1 - \xi_i, \ \forall i : y^i = +1 \\ & \left\langle w, x^i \right\rangle \leq b - 1 + \xi_i, \ \forall i : y^i = -1. \end{aligned}$$

- Note that in this formulation, it is evident that whenever $y^i = +1$, and if x^i lies on the improper side of the separating hyperplane, then the point x^i lies on a hyperplane of the form $\left\langle w, x^i \right\rangle = b + 1 - \xi_i$ for some $\xi^i > 0$. Note that this hyperplane is parallel to the separating and the supporting hyperplanes.
- **Exercise:** Check the analogous case for a point with $y^i = -1$ and which has improper orientation.

Soft-margin Support Vector Machine

Note that another equivalent formulation for soft-margin SVM is:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \max(0, 1 - y^i(\langle w, x^i \rangle - b))$$

- This formulation is of the form $\text{Reg}(w) + C\ \text{Loss}(\mathcal{D}; w, b)$, where $\text{Reg}(w)$ denotes the regularizer term in model parameters w, and $\text{Loss}(\mathcal{D}; w, b)$ denotes an aggregate loss term over the data set $\mathcal{D}$, and the loss term is parameterized by w, b.
- Note that linear regression and logistic regression problems can also be written in the form of regularized loss minimization. **(Exercise!)**

Soft-margin Support Vector Machine

Consider the equivalent formulation for soft-margin SVM:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \max(0, 1 - y^i(\langle w, x^i \rangle - b))$$

Here the sample-wise loss is given by:

$$\text{Loss}((x^i, y^i); w, b) = \max(0, 1 - y^i(\langle w, x^i \rangle - b)).$$

Soft-margin Support Vector Machine

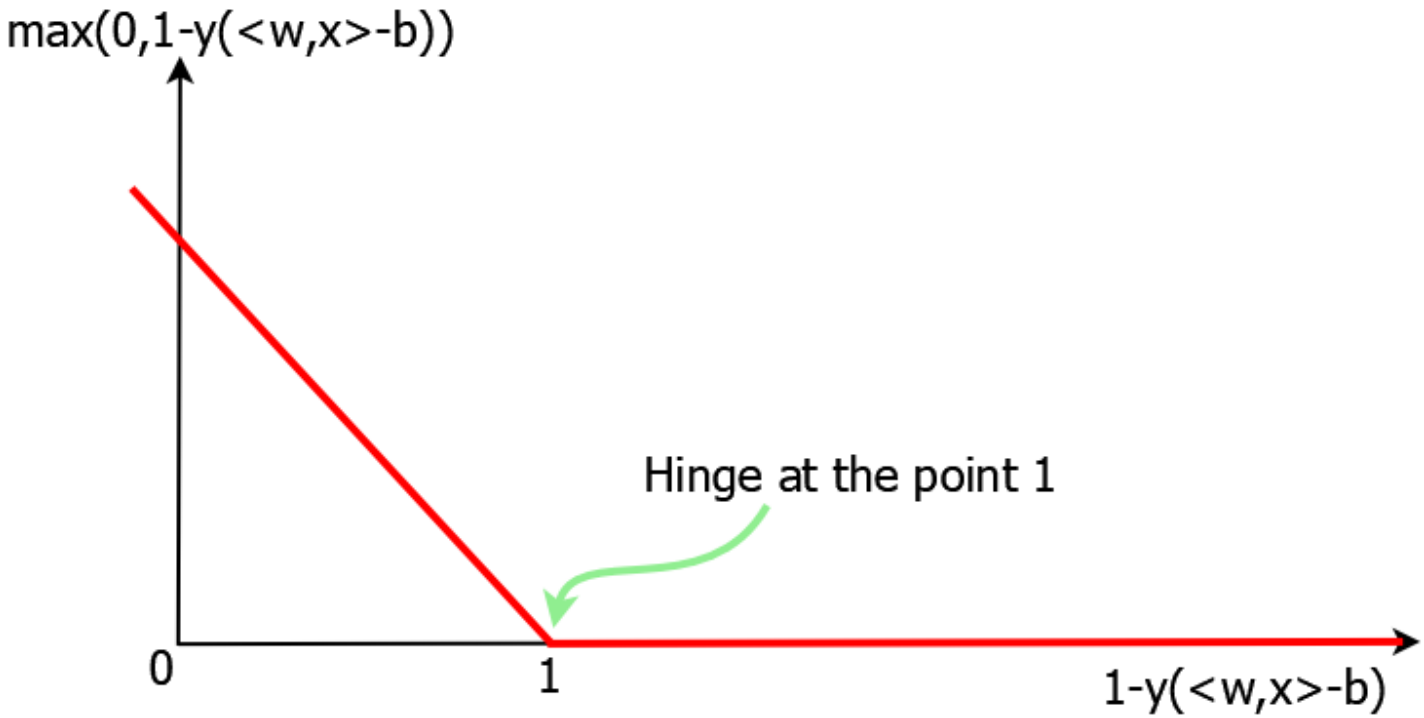


Figure: Hinge Loss

- Suppose we plot the quantity $1 - y^i(\langle w, x^i \rangle - b)$ on the horizontal axis against $\text{Loss}((x^i, y^i); w, b)$ on the vertical axis in a two-dimensional plot, we get a kink at the value 1.
- Since this resembles a hinge, the loss itself is called hinge loss.

Optimality conditions for constrained optimization problems

A generic constrained optimization problem

Consider an optimization problem written in a standard form:

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \ & f(w) \\ \text{s.t. } & g_i(w) \leq 0, \ \forall i \in \{1, 2, \ldots, p\} \\ & h_i(w) = 0, \ \forall i \in \{1, 2, \ldots, q\}. \end{aligned} \tag{OPT}$$

Note the following about problem (OPT):

- w denotes the decision or optimization variable which takes its values in $\mathbb{R}^d$.
- $f(w)$ denotes the objective function.
- The optimization objective is to minimize $f(w)$.
- There are p inequality constraints, and each inequality constraint is written with a ≤ 0 (less than or equal to 0) sense.
- There are q equality constraints.

A generic constrained optimization problem

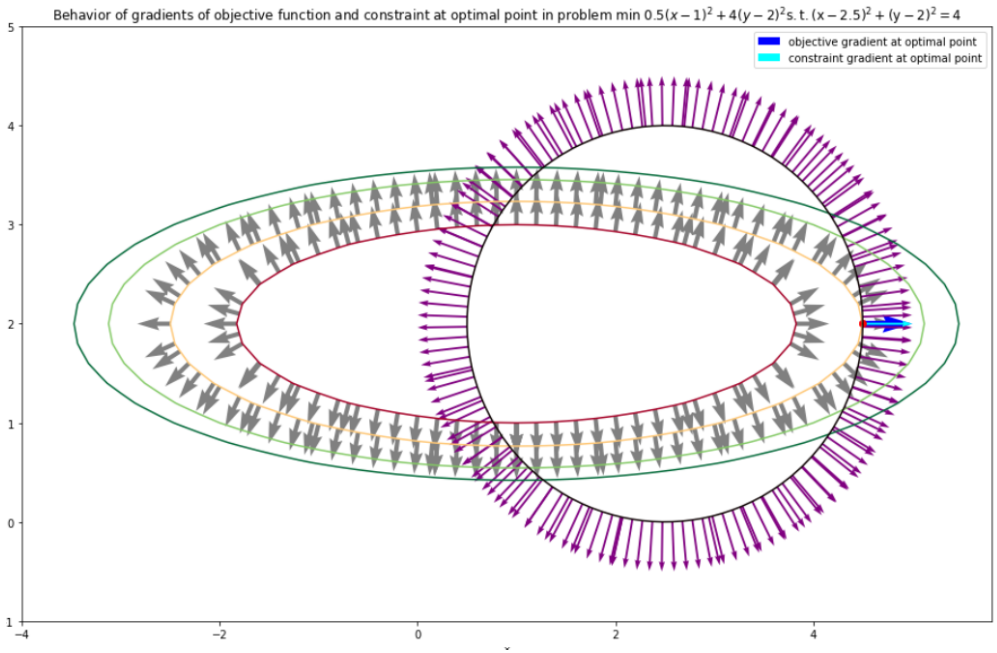
Consider an optimization problem written in a standard form:

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \; & f(w) \\ \text{s.t. } & g_i(w) \leq 0, \; \forall i \in \{1, 2, \ldots, p\} \\ & h_i(w) = 0, \; \forall i \in \{p+1, p+2, \ldots, p+q\}. \end{aligned} \tag{OPT}$$

- Recall that for unconstrained optimization problem of the form $\min_{w \in \mathbb{R}^d} f(w)$ where the **zero-gradient condition** $\nabla_w f(w) = 0$ becomes a necessary condition for optimality. Further **if f is convex in w**, the **zero-gradient condition** becomes a **sufficient condition for optimality**.
- However for constrained problems, the zero-gradient condition needs to be updated to incorporate the impact of the presence of constraints.

Optimality conditions for constrained optimization problems

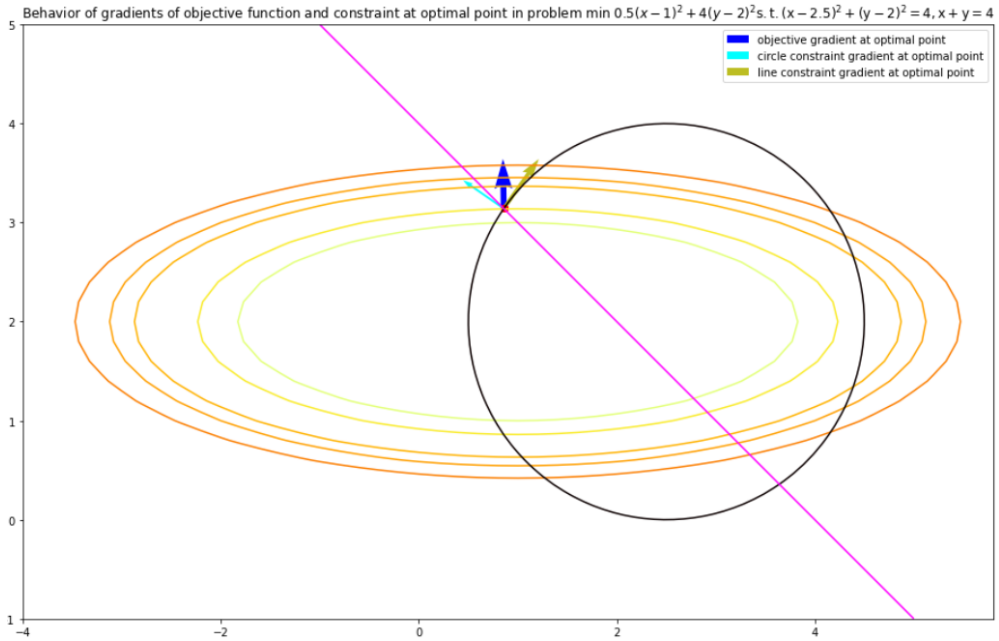
Let us see some examples of behavior of objective function and constraint gradients at optimal solution:



- In this figure, it is seen that the gradient of the objective function (in indigo color) and the gradient of the constraint function (in cyan color) are aligned in the same direction.

Optimality conditions for constrained optimization problems

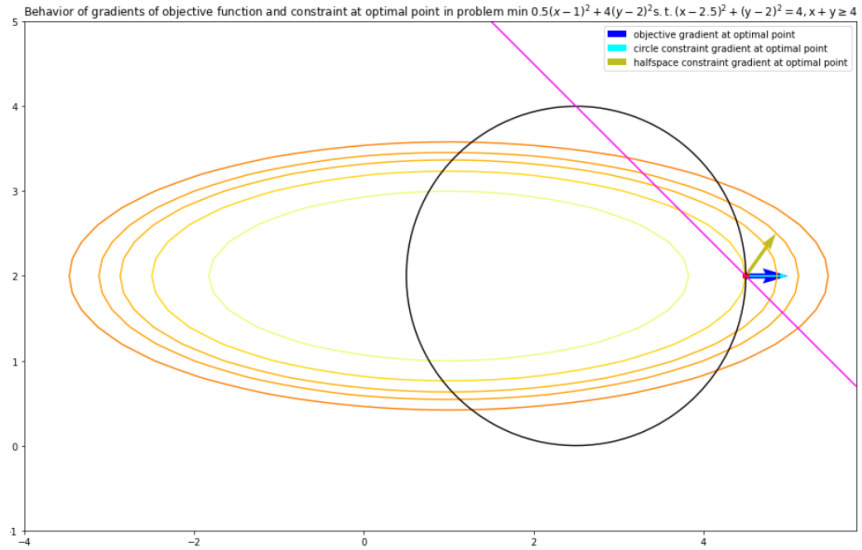
Let us see some examples of behavior of objective function and constraint gradients at optimal solution:



- In this figure, it is seen that the gradient of the objective function (in indigo color) lies in the span of gradient of the circle constraint (in cyan color) and gradient of the hyperplane equality constraint (in green color).

Optimality conditions for constrained optimization problems

Let us see some examples of behavior of objective function and constraint gradients at optimal solution:



- In this figure, it is seen that the gradient of the objective function (in indigo color) aligns with the gradient of the circle constraint (in cyan color) and is not dependent on the gradient of the hyperplane constraint (in green color).
- However it can be still seen that the gradient of objective function lies in the span of the gradients of the circle constraint and the hyperplane inequality constraint.

Optimality conditions for constrained optimization problems

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \ & f(w) \\ \text{s.t. } & g_i(w) \leq 0, \ \forall i \in \{1, 2, \ldots, p\} \\ & h_i(w) = 0, \ \forall i \in \{1, 2, \ldots, q\}. \end{aligned} \tag{OPT}$$

From previous observations, it can be said that for problem (OPT), the gradient $\nabla_w f(w)$ of objective function and the gradients $\nabla_w g_i(w)$ of inequality constraints and the gradients $\nabla_w h_i(w)$ of equality constraints satisfy the following relation:

$$\nabla_w f(w^*) = \sum_{i=1}^{p} \alpha_i \nabla_w g_i(w^*) + \sum_{i=1}^{q} \beta_i \nabla_w h_i(w^*) \tag{2}$$

for some suitable $\{\alpha_i\}_{i=1}^p, \{\beta_i\}_{i=1}^q$ at an optimal point w^*.

Optimality conditions for constrained optimization problems

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \ & f(w) \\ \text{s.t. } & g_i(w) \leq 0, \ \forall i \in \{1, 2, \ldots, p\} \\ & h_i(w) = 0, \ \forall i \in \{1, 2, \ldots, q\}. \end{aligned} \tag{OPT}$$

One candidate which gives explicit access to the following gradient span conditions of the form:

$$\nabla_w f(w^*) = \sum_{i=1}^{p} \alpha_i \nabla_w g_i(w^*) + \sum_{i=1}^{q} \beta_i \nabla_w h_i(w^*)$$

is the **Lagrangian** function of the optimization problem:

$$\mathcal{L}(w, \lambda, \mu) = f(w) + \sum_{i=1}^{p} \lambda_i g_i(w) + \sum_{i=1}^{q} \mu_i h_i(w), \tag{3}$$

Optimality conditions for constrained optimization problems

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \ & f(w) \\ \text{s.t. } & g_i(w) \leq 0, \ \forall i \in \{1, 2, \ldots, p\} \\ & h_i(w) = 0, \ \forall i \in \{1, 2, \ldots, q\}. \end{aligned} \tag{OPT}$$

The **Lagrangian** function of the optimization problem (OPT) is:

$$\mathcal{L}(w, \lambda, \mu) = f(w) + \sum_{i=1}^{p} \lambda_i g_i(w) + \sum_{i=1}^{q} \mu_i h_i(w).$$

The scalars $\{\lambda_i\}_{i=1}^{p}, \{\mu_i\}_{i=1}^{q}$ are called **Lagrange multipliers**.

Optimality conditions for constrained optimization problems

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \ & f(w) \\ \text{s.t. } & g_i(w) \leq 0, \ \forall i \in \{1, 2, \ldots, p\} \\ & h_i(w) = 0, \ \forall i \in \{1, 2, \ldots, q\}. \end{aligned} \tag{OPT}$$

Thus to access to the following gradient span conditions of the form:

$$\nabla_w f(w^*) = \sum_{i=1}^{p} \alpha_i \nabla_w g_i(w^*) + \sum_{i=1}^{q} \beta_i \nabla_w h_i(w^*)$$

we need to solve $\nabla_w \mathcal{L}(w, \lambda, \mu) = 0$.

Optimality conditions for constrained optimization problems

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \ & f(w) \\ \text{s.t. } & g_i(w) \leq 0, \ \forall i \in \{1, 2, \ldots, p\} \\ & h_i(w) = 0, \ \forall i \in \{1, 2, \ldots, q\}. \end{aligned} \tag{OPT}$$

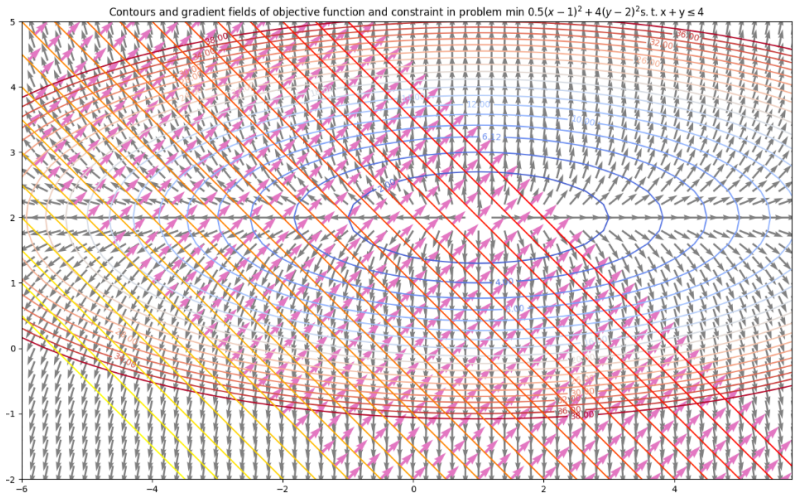
Thus to access to the following gradient span conditions of the form:

$$\nabla_w f(w^*) = \sum_{i=1}^{p} \alpha_i \nabla_w g_i(w^*) + \sum_{i=1}^{q} \beta_i \nabla_w h_i(w^*)$$

we need to solve $\nabla_w \mathcal{L}(w, \lambda, \mu) = 0$.
In fact, we will see that $\nabla_w \mathcal{L}(w, \lambda, \mu) = 0$ is one of the necessary conditions in constrained optimization.

Optimality conditions for constrained optimization problems

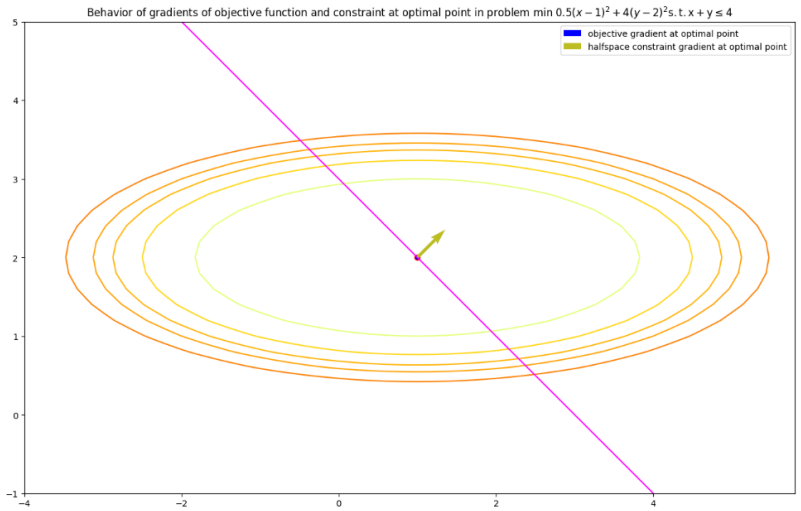
Let us see some examples of behavior of objective function and constraint gradient at optimal solution particularly for an optimization problem with a single inequality constraint $\min_{w \in \mathbb{R}^d} \ f(w)$ s.t. $g(w) \leq 0$:



- In this figure, from the contours of the objective function and the inequality constraint, we can see that the optimal solution lies within (or in the interior of) the constraint region.

Optimality conditions for constrained optimization problems

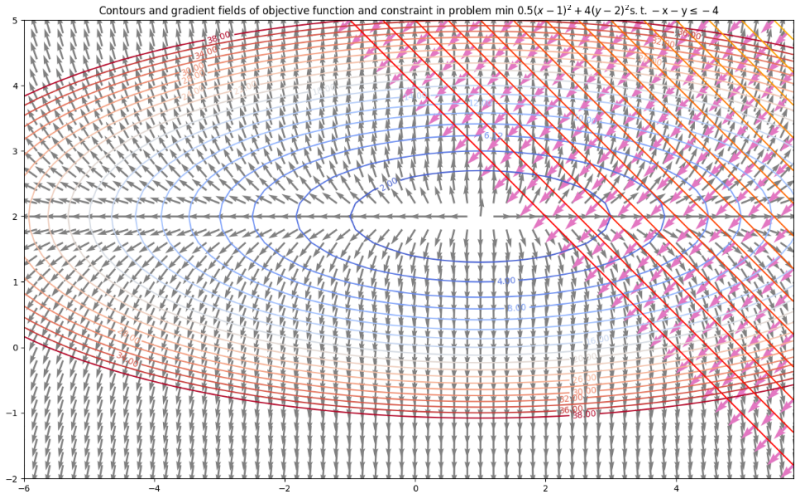
For the problem $\min_{w \in \mathbb{R}^d} f(w)$ s.t. $g(w) \leq 0$, let us see the behavior of gradients of objective function and the inequality constraint at optimal solution:



- In this figure, it is seen that the objective gradient of the optimal solution has a behavior similar to the behavior at an unconstrained solution. That is, the objective gradient is simply the zero vector (indicated as a dot in indigo color). Whereas, the gradient of the inequality constraint (in green color) seems to not determine the behavior of the objective gradient.
- Hence at that solution, $g(w^*) < 0$ and the corresponding Lagrange multiplier is zero.

Optimality conditions for constrained optimization problems

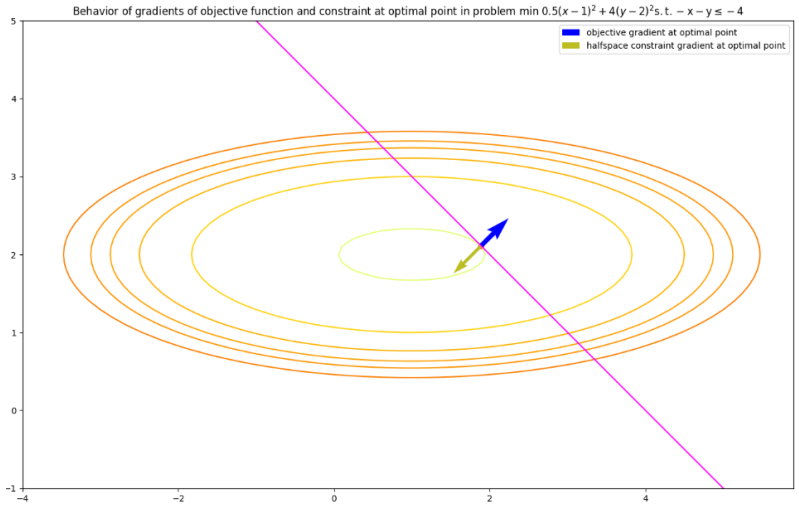
Let us see some examples of behavior of objective function and constraint gradient at optimal solution particularly for the problem $\min_{w \in \mathbb{R}^d} \; f(w)$ s.t. $g(w) \leq 0$:



- In this figure, from the contours of the objective function and the inequality constraint, we can see that the optimal solution lies on the boundary of the constraint region.

Optimality conditions for constrained optimization problems

For the previous example, let us see the behavior of gradients of objective function and the inequality constraint at optimal solution:



- From this figure, we note that when the gradient of the objective function and the gradient of the inequality constraint are aligned in opposite direction, the solution w^* necessarily lies on the boundary of the inequality constraint. Hence at that solution, $g(w^*) = 0$ holds.
- Further note that the corresponding Lagrange multiplier is strictly positive.

Optimality conditions for constrained optimization problems

From the previous examples we see that for a single inequality constraint, there are only two possibilities:

- The Lagrange multiplier associated with the inequality constraint is zero (or)
- The Lagrange multiplier associated with the inequality constraint is strictly positive.

Optimality conditions for constrained optimization problems

From the previous examples we see that for a single inequality constraint, there are only two possibilities:

- The Lagrange multiplier associated with the inequality constraint is zero (or)
- The Lagrange multiplier associated with the inequality constraint is strictly positive.

In fact, these observations extend to the general case for optimization problems with multiple inequality constraints and equality constraints.

Optimality conditions for constrained optimization problems

Combining these observations, we get the next set of optimality conditions:

- $\lambda_i = 0$ whenever $g_i(w^*) < 0$ and
- $g_i(w^*) = 0$ whenever $\lambda_i > 0$.
- These two conditions can be combined as the following **complementary slackness (CS) condition**:

$$\lambda_i g_i(w^*) = 0, \forall\ i \in \{1, 2, \ldots, p\}. \tag{4}$$

Optimality conditions for constrained optimization problems

These discussions lead to the following well-known Karush Kuhn Tucker (KKT) necessary conditions of optimality of constrained optimization problems at an optimal w^* and the corresponding Lagrange multipliers $\{\lambda_i^*\}_{i=1}^p, \{\mu_i^*\}_{i=1}^q$ (a.k.a. dual variables):

- Zero-gradient of Lagrangian w.r.t. primal variable:

$$\nabla_w \mathcal{L}(w^*, \lambda^*, \mu^*) = 0$$

$$\implies \nabla_w f(w^*) + \sum_{i=1}^{p} \lambda_i g_i(w^*) + \sum_{i=1}^{q} \mu_i h_i(w^*) = 0.$$

- Primal feasibility:
  - $g_i(w^*) \leq 0, \ \forall i \in \{1, 2, \ldots, p\}$.
  - $h_i(w^*) = 0, \ \forall i \in \{1, 2, \ldots, q\}$.
- Dual feasibility: $\lambda_i^* \geq 0, \ \forall i \in \{1, 2, \ldots, p\}$.
- CS condition: $\lambda_i^* g_i(w^*) = 0, \forall \ i \in \{1, 2, \ldots, p\}$.

Optimality conditions for constrained optimization problems

KKT Conditions:

- Zero-gradient of Lagrangian w.r.t. primal variable:

$$\nabla_w \mathcal{L}(w^*, \lambda^*, \mu^*) = 0$$

$$\implies \nabla_w f(w^*) + \sum_{i=1}^{p} \lambda_i g_i(w^*) + \sum_{i=1}^{q} \mu_i h_i(w^*) = 0.$$

- Primal feasibility:
  - $g_i(w^*) \leq 0, \ \forall i \in \{1, 2, \ldots, p\}$.
  - $h_i(w^*) = 0, \ \forall i \in \{1, 2, \ldots, q\}$.
- Dual feasibility: $\lambda_i^* \geq 0, \ \forall i \in \{1, 2, \ldots, p\}$.
- CS condition: $\lambda_i^* g_i(w^*) = 0, \forall \ i \in \{1, 2, \ldots, p\}$.

Note that the dual feasbility and CS conditions are related only to the inequality constraints.

Dual Function and Dual Problem

Dual function

- Note that the Lagrangian function of the optimization problem results in additional variables called Lagrange multipliers or dual variables.
- Based on the Lagrangian function, we construct a function that is dependent only on the dual variables. This function is called a **dual function** given by:

$$\mathscr{D}(\lambda, \mu) = \inf_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \mu). \qquad \text{(dual function)}$$

Dual function

- Note that the Lagrangian function of the optimization problem results in additional variables called Lagrange multipliers or dual variables.
- Based on the Lagrangian function, we construct a function that is dependent only on the dual variables. This function is called a **dual function** given by:

$$\mathscr{D}(\lambda, \mu) = \inf_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \mu). \qquad \text{(dual function)}$$

- Note that the dual function is a pointwise infimum over $w \in \mathbb{R}^d$.

Dual problem

- Based on the dual function, we define a different optimization problem which is characterized based on the Lagrangian multiplier variables.
- We call this problem dual problem.

$$\max_{\lambda \geq 0, \mu} \mathscr{D}(\lambda, \mu) \iff \max_{\lambda \geq 0, \mu} \inf_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \mu). \qquad \text{(dual problem)}$$

Dual problem

- Based on the dual function, we define a different optimization problem which is characterized based on the Lagrangian multiplier variables.
- We call this problem dual problem.

$$\max_{\lambda \geq 0, \mu} \mathscr{D}(\lambda, \mu) \iff \max_{\lambda \geq 0, \mu} \inf_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \mu). \qquad \text{(dual problem)}$$

- Based on the dual function and dual problem, we call the Lagrangian variables **dual variables**.

Dual problem

- Based on the dual function, we define a different optimization problem which is characterized based on the Lagrangian multiplier variables.
- We call this problem dual problem.

$$\max_{\lambda \geq 0, \mu} \mathscr{D}(\lambda, \mu) \iff \max_{\lambda \geq 0, \mu} \inf_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \mu). \qquad \text{(dual problem)}$$

- Based on the dual function and dual problem, we call the Lagrangian variables **dual variables**.
- Further the original variable w is called **primal variable**.

Dual problem

From the formulation of dual problem.

$$\max_{\lambda \geq 0, \mu} \mathscr{D}(\lambda, \mu) \iff \max_{\lambda \geq 0, \mu} \inf_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \mu).$$

we see a lower bound given by $\inf_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \mu)$.

Further note that the objective is to maximize the lower bound, where the maximization is with respect to the Lagrangian multiplier variables (or) dual variables.

SVM: Dual Problem Formulation

Dual problem of SVM

Let us now derive the dual problem of SVM.
Recall the (primal) soft-margin SVM optimization problem:

$$\min_{w,b,\xi_i \geq 0} \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{s.t. } y^i(\langle w, x^i \rangle - b) \geq 1 - \xi_i, \ \forall i \in \{1, 2, \ldots, n\}.$$

Dual problem of SVM

Writing the SVM optimization problem in standard form we have:

$$\min_{w,b,\xi_i} \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t. } 1 - \xi_i - y^i(\langle w, x^i\rangle - b) \leq 0, \ \forall i \in \{1,2,\ldots,n\},$$
$$-\xi_i \leq 0, \ \forall i \in \{1,2,\ldots,n\}.$$

Dual problem of SVM

The Lagrangian function of SVM problem is given by:

$$\mathcal{L}(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i[1 - \xi_i - y^i(\langle w, x^i\rangle - b)] + \sum_{i=1}^{n}\beta_i(-\xi_i)$$

Dual problem of SVM

To derive the SVM dual problem, we first use the KKT conditions for the primal SVM problem.

Dual problem of SVM

We have the following KKT conditions for primal SVM problem:

- Zero gradient condition of Lagrangian w.r.t. primal variables:

$$\nabla_w \mathcal{L} = 0, \nabla_b \mathcal{L} = 0,$$
$$\nabla_{\xi_i} \mathcal{L} = 0, \ \forall i \in \{1, 2, \ldots, n\}.$$

- Primal feasibility:

$$1 - \xi_i - y^i(\langle w, x^i \rangle - b) \leq 0, \ \forall i \in \{1, 2, \ldots, n\},$$
$$-\xi_i \leq 0, \ \forall i \in \{1, 2, \ldots, n\},$$

- Dual feasbility: $\alpha_i \geq 0, \beta_i \geq 0 \ \forall i \in \{1, 2, \ldots, n\}$.
- CS condition:
  - $\alpha_i[1 - \xi_i - y^i(\langle w, x^i \rangle - b)] = 0, \ \forall i \in \{1, 2, \ldots, n\}$.
  - $\beta_i \xi_i = 0 \ \forall \ i \in \{1, 2, \ldots, n\}$.

Dual problem of SVM

From the zero gradient conditions, we can derive:

$$\nabla_w \mathcal{L} = 0 \implies w = \sum_{i=1}^{n} \alpha_i y^i x^i,$$

$$\nabla_b \mathcal{L} = 0 \implies \sum_{i=1}^{n} \alpha_i y^i = 0$$

$$\nabla_{\xi_i} \mathcal{L} = 0 \implies C - \alpha_i - \beta_i = 0, \ \forall i \in \{1, 2, \ldots, n\}.$$

$$\implies 0 \leq \alpha_i \leq C \text{ (from dual feasbility conditions).}$$

Dual problem of SVM

From the zero gradient conditions, we can derive:

$$\nabla_w \mathcal{L} = 0 \implies w = \sum_{i=1}^{n} \alpha_i y^i x^i,$$

$$\nabla_b \mathcal{L} = 0 \implies \sum_{i=1}^{n} \alpha_i y^i = 0$$

$$\nabla_{\xi_i} \mathcal{L} = 0 \implies C - \alpha_i - \beta_i = 0, \ \forall i \in \{1, 2, \ldots, n\}.$$

$$\implies 0 \leq \alpha_i \leq C \text{ (from dual feasbility conditions).}$$

Note that these conditions are also sufficient for the inner problem $\inf_{w,b,\xi} \mathcal{L}(w, b, \xi, \alpha, \beta)$ involved in the dual function.

Dual problem of SVM

Substituting these into the Lagrangian, and after evaluating, we get the following dual problem for SVM:

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j (x^i)^\top (x^j) + \sum_{i=1}^{n}\alpha_i,$$

$$\text{s.t.} \sum_{i=1}^{n}\alpha_i y^i = 0,$$

$$0 \leq \alpha_i \leq C \ \forall i \in \{1, 2, \ldots, n\}. \qquad \text{(dual SVM)}$$

Exercise: Derive all details for obtaining dual SVM problem.

Inference for an unseen test sample

If the label for a test sample $\hat{x} \in \mathbb{R}^d$ is to be found, the prediction rule can be designed based on the appropriate half-space of the separating hyeprplane in which the test sample lies.

Exercise: Write a suitable inference rule for predicting the label for a test sample $\hat{x} \in \mathbb{R}^d$.

References:

- An Introduction to Support Vector Machines and Other Kernel-based Learning Methods by Nello Cristianini and John Shawe Taylor, Cambridge University Press.
- Chapter 7 in Pattern Recognition and Machine Learning by Chris Bishop (Springer publication).