Computational Data Analysis

Machine Learning

Yao Xie, Ph.D.

Associate Professor

Harold R. and Mary Anne Nash Early Career Professor

H. Milton Stewart School of Industrial and Systems Engineering

Basic Optimization

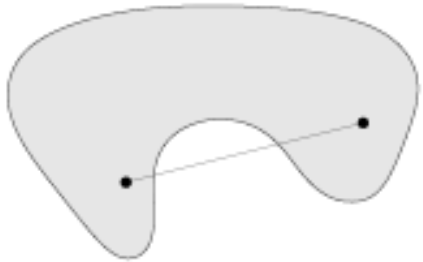
Outline

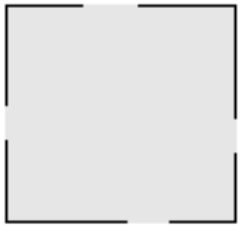
- Convex set
- Convex/cancave function
- First-order condition
- Second-order condition
- Theory under EM for GMM
- Langragian and dual function
- KKT conditions

The goal of this lecture is to build up necessary math background in developing algorithms for machine learning algorithms (previous and future lectures). You are not required to *master* everything.

Convex Set

- Definition: A set A is convex, if for every $0 \leq \alpha \leq 1$ it satisfies
  - $\forall x,\ y \in A \rightarrow \alpha x + (1-\alpha)y \in A$
- The line segment between any two points is also in the set.
- Examples of convex and non-convex sets





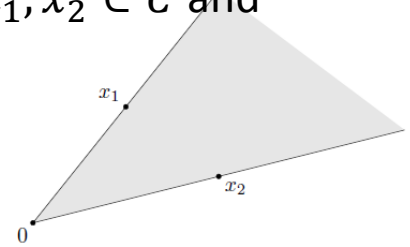
a set A is convex.
If for every alpha between 0 and
1, we can find any point, x and y belong to the set.
And linear combination in between is two points is still going to be with
belong to this set, okay.
So basically, the interpretation of this condition is that any line segment between any two points in the set, is also going to be in the set.
So if you look at the following example, the first example, is a convex set.

Common Convex Set

So a set C is a convex cone, if for any points within the set and if I just stretch this sub points by fraction theta 1 and 2. And they form a linear combination this is two within the sets. So these two fractions are going to be positive theta 1 and then theta 2. So conceptually, you can think about this set as some cone that defined passing through the origin. And so if i take any two points in my cone. I form a linear combination in the positive direction. Then the new combination between these two points is still. You're going to be between this two sets. So it's a cone because these two coefficients you can see theta1 and theta2 are going to be positive. So I cannot choose theta1 to be negative or theta2 to be negative. So this call is only pointed to is to one direction, the positive direction. That's why it's visually geometrically look like a cone.

- Cones: A set C is a convex cone, if for any $x_1, x_2 \in C$ and $\theta_1, \theta_2 \geq 0$, we have

$$\theta_1 x_1 + \theta_2 x_2 \in C$$



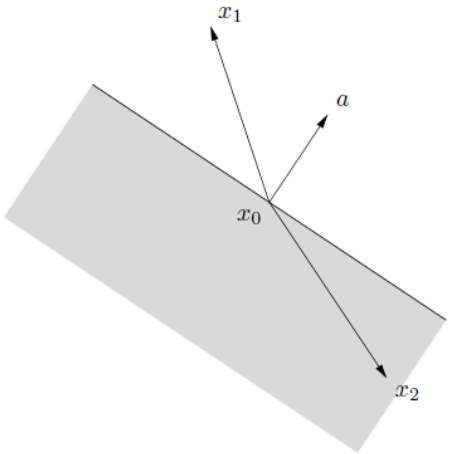
- Hyperplanes and halfspaces:

A set is hyperplane if

$$\{x | a^\top (x - x_0) = 0, a \neq 0\}$$

A halfspace is

$$\{x | a^\top (x - x_0) \leq 0, a \neq 0\}$$



Common Convex Set

The next example is ellipsoids the ellipsoids is a shape that's different from the ball in the sense you can have different Stretch in different directions.
Okay, so it's basically like a lips in high dimensional space.
And so for example, the definition is given by this E, that's going to be all the points such that x minus xc transpose times is P inverse, the T is a matrix, and times x minus xc less than equal to one.
And you can see this definition of the ellipse, first is the ellipsoids is gonna be centered around a point, axis E.
And the second important quantity of the parameter is p inverse.
This is the covariance matrix of the points.
And basically specifies the orientation and described raising issue of the axis for the xy definition.

- Euclidean balls: A Euclidean ball has the form

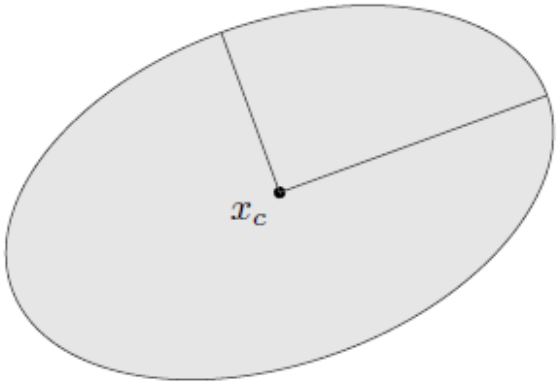
$$B(x_c, r) = \{x | \|x - x_c\|_2 \leq r\}$$

And this ball definition is defined in here.
The b denotes ball, and xc is the center of the ball, and r is radius of the ball.
As you can see this Convex sets, is prioritized by two parameters,
the center and the radius.
And this is basically everything centered around this point xc and

- Ellipsoids:

$$E = \{x | (x - x_c)^\top P^{-1} (x - x_c) \leq 1\}$$

- The eigen-vectors and eigen-values determine the direction and shape of the semi-axes



Used in support vector novelty detection

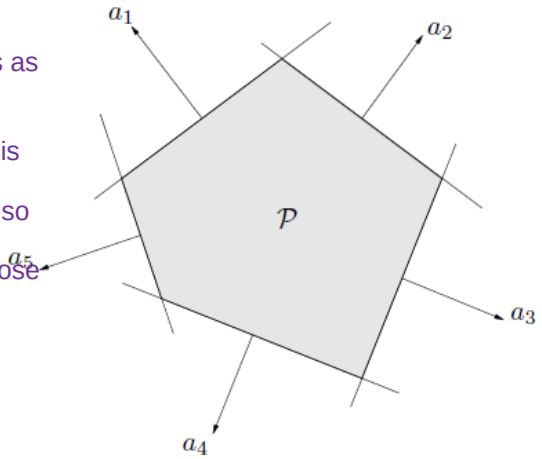
Common Convex Set

- Polyhedra: Intersection of a *finite* set of halfspaces/hyperplanes

$$P = \{x | a_j^\top x \leq b_j, j = 1, \ldots, m, c_j^\top x = d_j, j = 1, \ldots, p\}$$

- It is defined by as the solution set of a finite number of linear equalities and inequalities

The next example is polyhedra.
And this polyhedra is intersection of a finite set of halfspace or hyperplanes.
And as you can see, mathematically I can define this as instead of just
one definition of equalities.
Now, I'm gonna have m such inequalities, if equality is represented by
AJ transpose x less equal to BJ, and then you can also have
some economic constraints in the form of CG transpose x is equal to DJ.
And so, geometrically this defines a sad as you can see, I joined a bunch of planes
and the intersection on this plane defines the boundaries of my set.
And I consider everything that lies within
this intersection which is defined my convex sets.



Used in SVM

Operations that Preserve Convex Sets

- Intersections: In fact, *every* closed convex set S is the intersection of all halfspaces that contain it:

$$S = \bigcap \{H | H \text{ is halfspace}, S \subset H\}$$

- Linear combination:

$$\alpha S = \{\alpha x | x \in S\},\ S + \alpha = \{x + a | x \in S\}$$

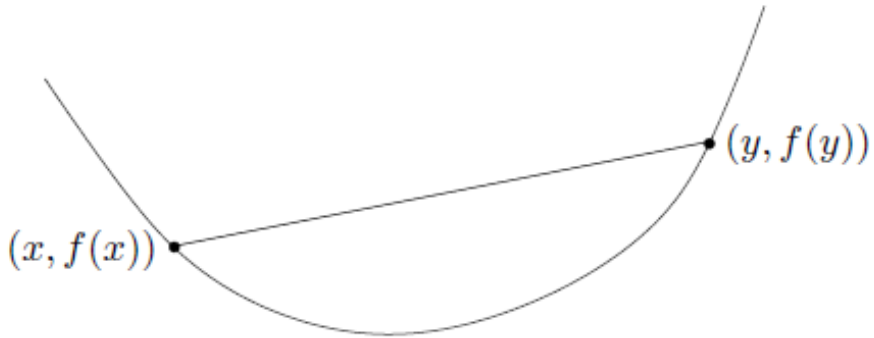
- Projection/Concatenation

Convex Functions

- Definition: A function $f: R^n \rightarrow R$ is convex if the domain $\mathbf{dom} f$ is a convex set and if for all $x, y \in \mathbf{dom} f$, and $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

- Geometrically, the line segment between $(x, f(x))$ and $(y, f(y))$ lies above the graph of f

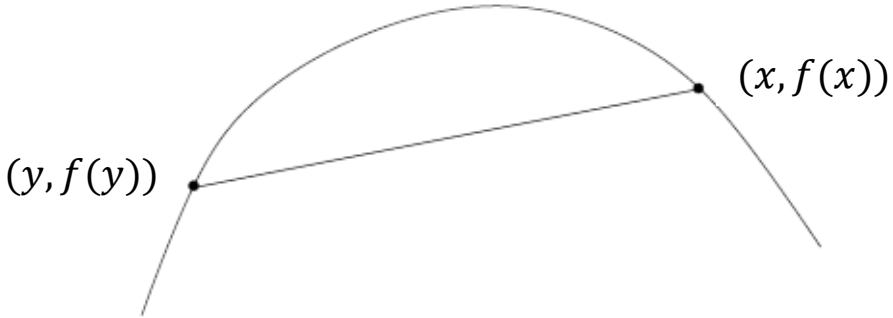


Concave Functions

- Definition: A function $f: R^n \rightarrow R$ is concave if the domain $\mathbf{dom} f$ is a convex set and if for all $x, y \in \mathbf{dom} f$, and $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta) f(y)$$

- Geometrically, the line segment between $(x, f(x))$ and $(y, f(y))$ lies below the graph of f

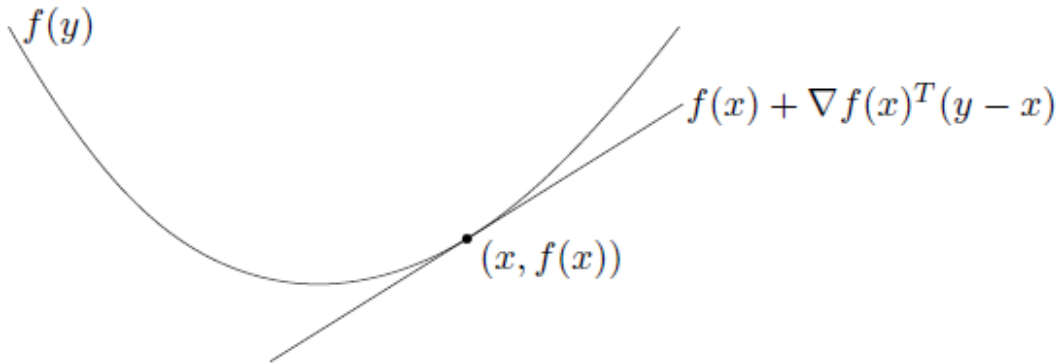


First-order Condition

And another way to characterize convexity is through the so-called first order condition.

- If f is differentiable, another way to characterize it is the first-order condition: f is convex iff $\mathbf{dom} f$ is convex and
$$f(y) \geq f(x) + \nabla f(x)^{\mathsf{T}}(y - x)$$
holds for all $x, y \in \mathbf{dom} f$.

- Geometrically, it means that the tangent line of f at point x lies below the function



unconstrained optimization

If f is a differentiable, you take a function that you can take derivative.
And another way to characterize
convexity is through the so-called first order condition.
And so this means the following, f is convex,
if and only if the first domain of f have to become x.
And the second is f(y) is greater than equal to f(x) + the gradient vector of f(x).
So since f is multivariate, so it's gradient is going to be a vector.
So this factor in your product with y- x holds for any x and y in my domain.
And geometrically what this means is, if I have a function that is convex,
then I take any points f(x) on my function.
And now I draw attention to my parts of this point.
Then tangent line is always going to lie below my function.
Again, this is very intuitive.
As you can see if you function shape is convex, then it should
be like about a linear approximation of your function at any given point.

Second-order Condition

- If f is twice differential, the second-order condition is: f is convex iff $\mathbf{dom} f$ is convex and for all $x \in \mathbf{dom} f$

  $$\nabla^2 f(x) \succcurlyeq 0$$

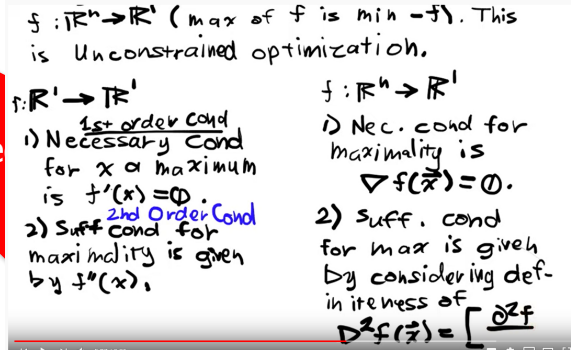
  positive semidefinite (symmetric and all eigenvalue nonnegative)

- That is the Hessian is positive semidefinite.

- Geometrically, the graph of the function has positive (upward) curvature at every point.

- Eg. $\mathrm{f}(\mathrm{x}) = x^{\mathrm{T}} A x$, for A positive semidefinite

Use



And by the way, what does it mean for a matrix to be positive semidefinite?
It's basically any matrix that is symmetric.
And if you look at the eigenvalues of this matrix, it has to be nonnegative, okay?
So this is also called the summation is positive semidefinite.
This matrix,
second-order derivative matrix is also called the Hermaissian matrix.

Examples

- Exponential: e^{ax} for every $a \in R$
- Powers: x^a is convex on R_{++} when $a \geq 1$ or $a \leq 0$; concave (i.e., $-f$ is convex) for $0 \leq a \leq 1$
- Powers of absolute vaue: $|x|^p$ for $p \geq 1$
- Logarithm: $\log x$ is concave on R_{++}
- Negative entropy: $x \log x$ is convex
- Norms: All norms are convex (nonnegative; homogeneous; triangular inequality)
- Max function: $f(x) = \max\{x_1, \ldots, x_n\}$ is convex
- Log-determinant: $f(X) = \log \det X$ is convex for all positive definite matrices

Used in EM

Used in multivariate Gaussian fit

Operations that Preserve Convexity

- Nonnegative weighted sums: If $f_1, \dots, f_m$ are convex, and $w_1, \dots, w_m \geq 0$, then
$$f = w_1 f_1 + \cdots + w_m f_m$$
is convex

- Composition with an affine mapping: suppose f is convex, then
$$g(x) = f(Ax + b)$$
with $\mathbf{dom} g = \{x | Ax + b \in \mathbf{dom} f\}$ is convex

- Pointwise maximum and supremum: If f_1 and f_2 are convex, then $f(x) = \max\{f_1, f_2\}$ is also convex. It easily extends to multiple functions.

Operations that Preserve Convexity

- Composition: If h is convex and nondecreasing, and g is convex, then $f(x) = h(g(x))$ is convex
  - The second derivative of f is
    $$f''(x) = h''\big(g(x)\big)g'(x)^2 + h'\big(g(x)\big)g''(x)$$
    for f to be convex, f'' should be nonnegative

- Log-sum-exp: $f(x) = \log(e^{x_1} + \cdots + e^{x_n})$

Theory underlying EM

- Recall that in MLE, we intend to learn the model parameter that would have maximize the likelihood of the data.
  - $l(\theta; D) = \log \sum_z p(x, z|\theta) = \log \sum_z p(x|\, z, \theta) P(z|\theta)$
- But we are iterating these:
  - Expectation step (E-step)
    - $f(\theta) = E_{q(z)}[\log p(x, z|\theta)], where\ q(z) = P(z|x, \theta^t)$
  - Maximization step (M-step)
    - $\theta^{t+1} = argmax_\theta\ f(\theta)$
- Does maximizing this surrogate yield a maximizer of the likelihood?

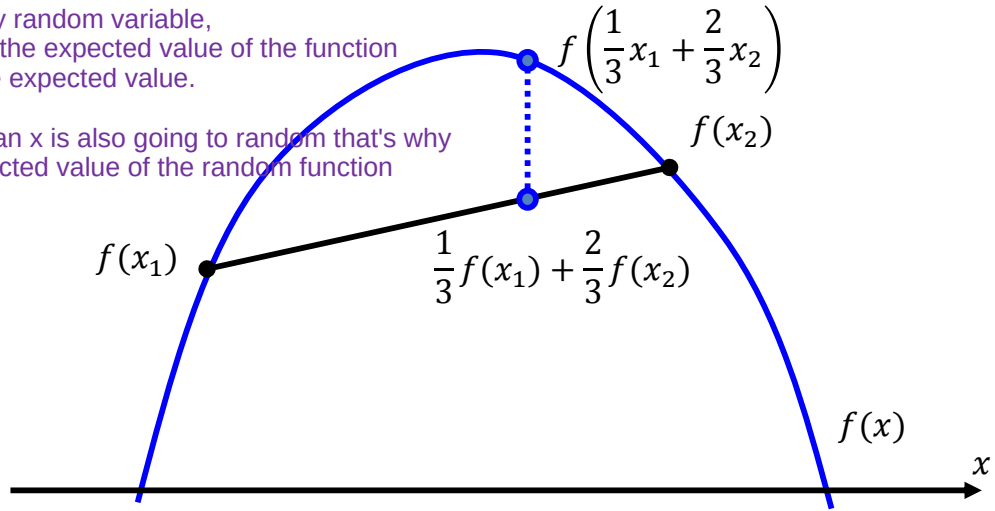
Jensen's inequality

- For concave function $f(x)$
  - $f(\sum_i \alpha_i x_i) \geq \sum_i \alpha_i f(x_i)\,, where\ \sum_i \alpha_i = 1, \alpha_i \geq 0$

- Most general case: If x is a random variable, and f is concave,

$$f(\mathbf{E}x) \geq \mathbf{E}f(x)$$

if I have x which is random variable.
Then if I look at extracting value of my random variable,
the function if it's concave, applied to the expected value of the function
should be greater than or equal to the expected value.
Of my f applied my x.
So since x is random, the f, applying an x is also going to random that's why
here we have an expectation as expected value of the random function



Lower bound of log-likelihood

So getting back to our IM problem our log-likelihood function is going to be
log of this sum of z the z is my latent factor indicating which component the sample comes from

- Log-likelihood $l(x;\theta) = \log \sum_z p(x,z|\theta)$

The log function is concave.

$$= \log \sum_z q(z) \frac{p(x,z|\theta)}{q(z)} \quad (\text{arbitrary } q(z))$$

This is a value, I can think about it as distribution function for the q.

$$\geq \sum_z q(z) \log \frac{p(x,z|\theta)}{q(z)} \quad (Jensen's\ inequality\ f\left(\sum_i \alpha_i x_i\right) \geq \sum_i \alpha_i f(x_i))$$

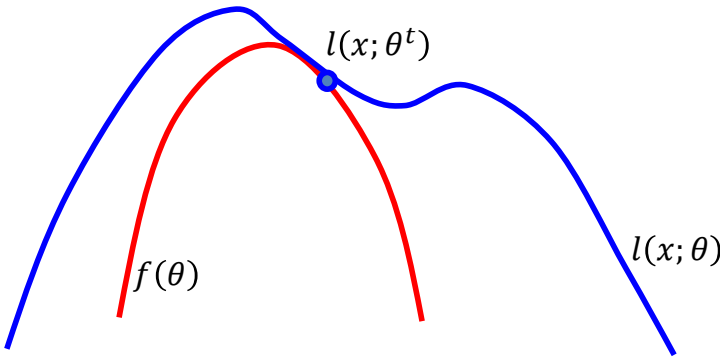
$$= \sum_z q(z) \log p(x,z|\theta) - \sum_z q(z) \log q(z)$$

$$= E_{q(z)}[\log p(x,z|\theta)] + H_{q(z)}$$

What q to use?

What attains equality?

- $q(z) = p(z|x, \theta^t)$: posterior of z given x attains the equality at θ^t
- Let $F(q, \theta) = \sum_z q(z|x) \log \frac{p(x,z|\theta)}{q(z|x)} \leq l(x; \theta) = \log \sum_z p(x, z|\theta)$
- $F(p(z|x, \theta^t), \theta^t) = \sum_z p(z|x, \theta^t) \log \frac{p(x,z|\theta^t)}{p(z|x,\theta^t)}$
- $= \sum_z p(z|x, \theta^t) \log p(x|\theta^t)$
- $= \log p(x|\theta^t)$
- $= \log \sum_z p(x, z|\theta^t)$



Convex Optimization

- Definition: An optimization problem is specified by

$$\begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && f_i(x) \leq 0, \ i = 1, \ldots, m \\ &&& h_i(x) = 0, \ i = 1, \ldots, p \end{aligned}$$

these are the constrained functions.

- A convex optimization problem has the following requirements
  - The objective function $f_0(x)$ must be convex
  - The inequality constraint functions $f_i(x)$ must be convex
  - The equality constraint functions $h_i(x)$ must be affine or linear

- Eg. support vector machines (SVM), logistic regression, maximum likelihood, ridge regression, …

Convex Optimization

- Global optimum: a point x^* in the feasible set is a global optimum iff
$$f_0(x^*) \leq f_0(x)$$
for all x in the feasible set

- Local optimum: a point x^* in the feasible set is a local optimum iff there exists $r > 0$, such that for all $x \in \{x | \|x - x^*\| \leq r\}$ and also in the feasible set, we have $f_0(x^*) \leq f_0(x)$

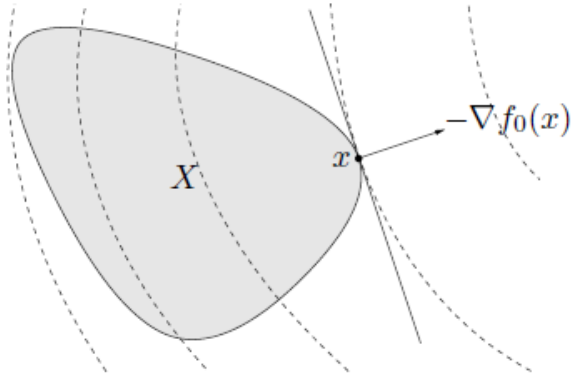
- For convex optimization problem, any local optimum is also a global optimum

First order optimality condition

- Let X denotes the feasible set, then x is optimal iff
$$\nabla f_0(x)^\mathsf{T}(y - x) \geq 0 \text{ for all } y \in X$$

- For an unconstrained problem, the condition becomes
$$\nabla f_0(x) = 0$$

- Geometrically, if $\nabla f_0(x) \neq 0$, it means $-\nabla f_0(x)$ is tangent to the feasible set at x



~~Operations that Preserve Convex Sets~~ Lagrangian

- For an unconstrained problem, the condition becomes

$$\nabla f_0(x) = 0$$

- For an constrained problem, we need to use the Lagrangian

$$L(x, \mu, \lambda) = f_0(x) + \sum_{i=1}^{p} \mu_i h_i(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$$

$$\text{s.t.} \quad \lambda_i \geq 0$$

to transform it into an unconstrained problem

Lagrangian will make the constrained problem to be treated as non-constrained problem.

- It is a lower bound of $f_0(x)$ for all $x \in X$, since $h_i(x)=0$, $f_i(x) \leq 0$ and $\lambda_i \geq 0$

$$L(x, \mu, \lambda) \leq f_0(x) \text{ for all } x \in X$$

Used in SVM

Lagrange dual function

- The Lagrange dual function is

$$g(\mu, \lambda) = \inf_x L(x, \mu, \lambda)$$

- It is a lower bound for the optimal value

$$g(\mu, \lambda) = \inf_x L(x, \mu, \lambda) \leq L(x^*, \mu, \lambda) \leq f_0(x^*)$$

- We want to maximize the lower bound to make it tight

$$g(\mu^*, \lambda^*) = \max g(\mu, \lambda)$$

Primal and Dual problems

I can either solve the primal or dual problem.
The result would be the same.

we solve the dual problem because it is easier.

- Primal problem

$$\begin{aligned} &\text{minimize} \quad f_0(x) \\ &\text{subject to} \quad f_i(x) \leq 0, \ i = 1, \dots, m \\ &\qquad\qquad h_i(x) = 0, \ i = 1, \dots, p \end{aligned}$$

- Dual problem

$$\begin{aligned} &\text{maximize} \quad g(\mu, \lambda) \\ &\text{subject to} \quad \lambda_i \geq 0, \ i = 1, \dots, m \end{aligned}$$

Used in SVM

solving the dual problem is to find the lower bound for primal problem.

- Strong duality (for convex problems, with mild condition)

$$g(\mu^*, \lambda^*) = f_0(x^*)$$

- Slater's condition: There exists an x inside the relative interior of the domain X such that, $f_i(x) < 0, \ i = 1, \dots, m$

KKT Optimality conditions

- The following list of optimality conditions, for an optimal triplet (x^*, μ^*, λ^*), are called KKT conditions

- $\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \mu_i^* \nabla h_i(x^*) = 0$

- $\lambda_i^* f_i(x^*) = 0$

- $f_i(x^*) \leq 0$

- $h_i(x^*) = 0$

- $\lambda_i^* \geq 0$


SUMMARY