Lecture 17 SVM (Part II) and Online Learning

Lei Li, **Yu-Xiang Wang**

(some slides from my convex optimization class,
originally taught by Ryan Tibshirani in CMU)

Recap: Support Vector Machines

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$ having rows $x_1, \ldots x_n$, recall the support vector machine or SVM problem:

$$
\begin{aligned}
\min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^n \xi_i \\
\text{subject to} \quad & \xi_i \geq 0, \; i = 1, \ldots n \\
& y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \; i = 1, \ldots n
\end{aligned}
$$

This is a quadratic program

Recap: Lagrange dual problem

Given a minimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0, \ i = 1, \dots m \\ & \ell_j(x) = 0, \ j = 1, \dots r \end{aligned}$$

we defined the Lagrangian:

$$L(x, u, v) = f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j \ell_j(x)$$

and Lagrange dual function:

$$g(u, v) = \min_x \ L(x, u, v)$$

Recap: Lagrange dual problem

The subsequent dual problem is:

$$\max_{u,v} \quad g(u,v)$$
$$\text{subject to} \quad u \geq 0$$

Important properties:

- Dual problem is always convex, i.e., g is always concave (even if primal problem is not convex)
- The primal and dual optimal values, $f^\star$ and $g^\star$, always satisfy weak duality: $f^\star \geq g^\star$
- Slater's condition: for convex primal, if there is an x such that

  $$h_1(x) < 0, \ldots h_m(x) < 0 \quad \text{and} \quad \ell_1(x) = 0, \ldots \ell_r(x) = 0$$

  then strong duality holds: $f^\star = g^\star$. Can be further refined to strict inequalities over the nonaffine h_i, $i = 1, \ldots m$

Recap: Deriving the dual of SVM

Introducing dual variables $v, w \geq 0$, we form the Lagrangian:

$$L(\beta, \beta_0, \xi, v, w) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n v_i\xi_i + \sum_{i=1}^n w_i\big(1 - \xi_i - y_i(x_i^T\beta + \beta_0)\big)$$

Recap: Dual SVM

Minimizing over β, β_0, ξ gives Lagrange dual function:

$$g(v, w) = \begin{cases} -\frac{1}{2} w^T \tilde{X} \tilde{X}^T w + 1^T w & \text{if } w = C1 - v,\ w^T y = 0 \\ -\infty & \text{otherwise} \end{cases}$$

where $\tilde{X} = \mathrm{diag}(y) X$. Thus SVM dual problem, eliminating slack variable v, becomes

$$\begin{aligned} \max_{w} \quad & -\frac{1}{2} w^T \tilde{X} \tilde{X}^T w + 1^T w \\ \text{subject to} \quad & 0 \le w \le C1,\ w^T y = 0 \end{aligned}$$

Check: Slater's condition is satisfied, and we have strong duality. Further, from study of SVMs, might recall that at optimality

$$\beta = \tilde{X}^T w$$

This is not a coincidence, as we'll later via the KKT conditions

“Kernel trick” in SVM

- The dual SVM depends only on inner products

$$\begin{array}{ll} \max_{w} & -\frac{1}{2}w^T \tilde{X}\tilde{X}^T w + 1^T w \\ \text{subject to} & 0 \le w \le C1,\ w^T y = 0 \end{array}$$

- How to make predictions?

This lecture

- KKT conditions
  - SVM as an example

- Online Learning

Optimality conditions: the conditions that characterizes the optimal solutions

- What you learned in high school

$$\min_{x \in \mathbb{R}} x^2 - 4x + 9$$

- Slight generalization: For convex and differentiable objective function

$$\min_{x \in \mathbb{R}^d} f(x)$$

Does not handle non-differentiable functions, does not handle constraints.

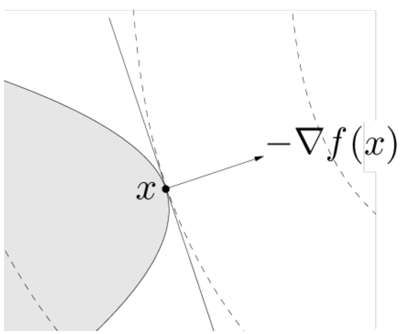
Handling constraints with first-order optimality conditions

For a convex problem

$$\min_x \; f(x) \quad \text{subject to} \quad x \in C$$

and differentiable f, a feasible point x is optimal if and only if

$$\nabla f(x)^T (y - x) \geq 0 \quad \text{for all } y \in C$$



This is called the first-order condition for optimality

In words: all feasible directions from x are aligned with gradient $\nabla f(x)$

Important special case: if $C = \mathbb{R}^n$ (unconstrained optimization), then optimality condition reduces to familiar $\nabla f(x) = 0$

Handling non-differentiable functions with **"subgradient"**

Recall that for convex and differentiable f,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y$$

I.e., linear approximation always underestimates f

A subgradient of a convex function f at x is any $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T (y - x) \quad \text{for all } y$$

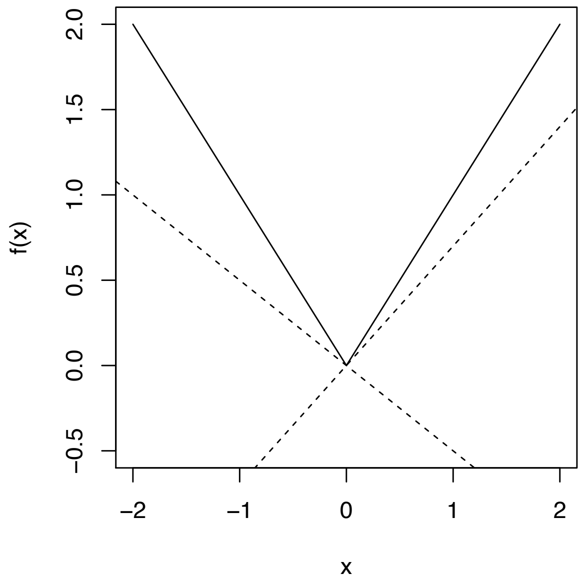
- Always exists[1]
- If f differentiable at x, then $g = \nabla f(x)$ uniquely
- Same definition works for nonconvex f (however, subgradients need not exist)

---

[1] On the relative interior of $\mathrm{dom}(f)$

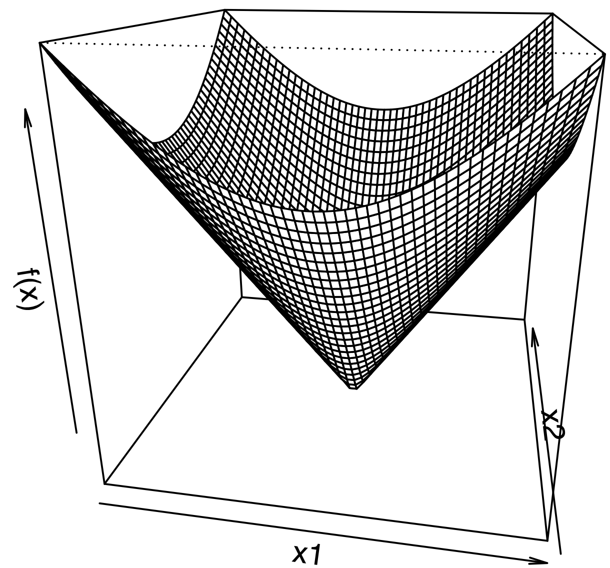
Examples of subgradients

Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$



- For $x \neq 0$, unique subgradient $g = \mathrm{sign}(x)$
- For $x = 0$, subgradient g is any element of $[-1, 1]$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2$



- For $x \neq 0$, unique subgradient $g = x/\|x\|_2$
- For $x = 0$, subgradient g is any element of $\{z : \|z\|_2 \leq 1\}$

Subdifferential

Set of all subgradients of convex f is called the subdifferential:

$$\partial f(x) = \{g \in \mathbb{R}^n : g \text{ is a subgradient of } f \text{ at } x\}$$

- Nonempty (only for convex f)
- $\partial f(x)$ is closed and convex (even for nonconvex f)
- If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- If $\partial f(x) = \{g\}$, then f is differentiable at x and $\nabla f(x) = g$

First order optimality condition with subgradient

For any f (convex or not),

$$f(x^\star) = \min_x f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(x^\star)$$

I.e., $x^\star$ is a minimizer if and only if 0 is a subgradient of f at $x^\star$.
This is called the subgradient optimality condition

Why? Easy: $g = 0$ being a subgradient means that for all y

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star)$$

Note the implication for a convex and differentiable function f, with $\partial f(x) = \{\nabla f(x)\}$

Karush-Kuhn-Tucker conditions

Given general problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0, \ i = 1, \dots m \\ & \ell_j(x) = 0, \ j = 1, \dots r \end{aligned}$$

The Karush-Kuhn-Tucker conditions or KKT conditions are:

- $0 \in \partial\Big(f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)\Big)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
- $h_i(x) \le 0, \ \ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \ge 0$ for all i (dual feasibility)

Necessity

Let $x^\star$ and $u^\star, v^\star$ be primal and dual solutions with zero duality gap (strong duality holds, e.g., under Slater's condition). Then

$$\begin{aligned} f(x^\star) &= g(u^\star, v^\star) \\ &= \min_x \; f(x) + \sum_{i=1}^m u_i^\star h_i(x) + \sum_{j=1}^r v_j^\star \ell_j(x) \\ &\le f(x^\star) + \sum_{i=1}^m u_i^\star h_i(x^\star) + \sum_{j=1}^r v_j^\star \ell_j(x^\star) \\ &\le f(x^\star) \end{aligned}$$

In other words, all these inequalities are actually equalities

Two things to learn from this:

- The point $x^\star$ minimizes $L(x, u^\star, v^\star)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^\star, v^\star)$ must contain 0 at $x = x^\star$—this is exactly the stationarity condition
- We must have $\sum_{i=1}^m u_i^\star h_i(x^\star) = 0$, and since each term here is ≤ 0, this implies $u_i^\star h_i(x^\star) = 0$ for every i—this is exactly complementary slackness

Primal and dual feasibility hold by virtue of optimality. Therefore:

If $x^\star$ and $u^\star, v^\star$ are primal and dual solutions, with zero duality gap, then $x^\star, u^\star, v^\star$ satisfy the KKT conditions

(Note that this statement assumes nothing a priori about convexity of our problem, i.e., of f, h_i, ℓ_j)

Sufficiency

If there exists $x^\star, u^\star, v^\star$ that satisfy the KKT conditions, then

$$\begin{aligned} g(u^\star, v^\star) &= f(x^\star) + \sum_{i=1}^{m} u_i^\star h_i(x^\star) + \sum_{j=1}^{r} v_j^\star \ell_j(x^\star) \\ &= f(x^\star) \end{aligned}$$

where the first equality holds from stationarity, and the second holds from complementary slackness

Therefore the duality gap is zero (and $x^\star$ and $u^\star, v^\star$ are primal and dual feasible) so $x^\star$ and $u^\star, v^\star$ are primal and dual optimal. Hence, we've shown:

If $x^\star$ and $u^\star, v^\star$ satisfy the KKT conditions, then $x^\star$ and $u^\star, v^\star$ are primal and dual solutions

Putting it together

In summary, KKT conditions:

- always sufficient
- necessary under strong duality

Putting it together:

For a problem with strong duality (e.g., assume Slater's condition: convex problem and there exists x strictly satisfying non-affine inequality contraints),

$$x^\star \text{ and } u^\star, v^\star \text{ are primal and dual solutions}$$
$$\iff x^\star \text{ and } u^\star, v^\star \text{ satisfy the KKT conditions}$$

(Warning, concerning the stationarity condition: for a differentiable function f, we cannot use $\partial f(x) = \{\nabla f(x)\}$ unless f is convex! There are other versions of KKT conditions that deal with local optima.)

Example: support vector machines

Given $y \in \{-1, 1\}^n$, and $X \in \mathbb{R}^{n \times p}$, the support vector machine problem is:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \ i = 1, \ldots n \\ & y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots n \end{aligned}$$

Introduce dual variables $v, w \geq 0$. KKT stationarity condition:

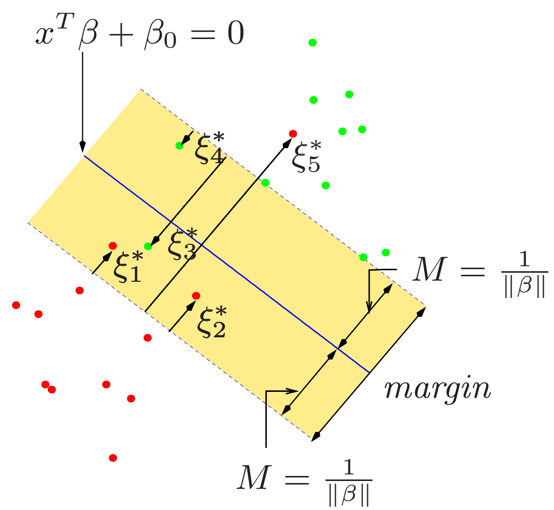
$$0 = \sum_{i=1}^n w_i y_i, \quad \beta = \sum_{i=1}^n w_i y_i x_i, \quad w = C1 - v$$

Complementary slackness:

$$v_i \xi_i = 0, \ w_i\big(1 - \xi_i - y_i(x_i^T \beta + \beta_0)\big) = 0, \quad i = 1, \ldots n$$

Hence at optimality we have $\beta = \sum_{i=1}^n w_i y_i x_i$, and w_i is nonzero only if $y_i(x_i^T \beta + \beta_0) = 1 - \xi_i$. Such points i are called the support points

- For support point i, if $\xi_i = 0$, then x_i lies on edge of margin, and $w_i \in (0, C]$;
- For support point i, if $\xi_i \neq 0$, then x_i lies on wrong side of margin, and $w_i = C$



KKT conditions do not really give us a way to find solution, but gives a better understanding

In fact, we can use this to screen away non-support points before performing optimization

Checkpoint: KKT conditions and SVM

- A generalized set of conditions that characterizes the optimal solutions
  - Stationarity, complementary slackness, primal / dual feasibility
  - Always sufficient for optimality
  - Necessary when we have strong duality

- Complementary slackness implies
  - SVM dual solutions are sparse!
  - The number of “support vector”s is small

This lecture

- KKT conditions
  - SVM as an example
- Online Learning

Recap: Statistical Learning Setting

(Adversarial) Online Learning Setting

- Data points show up sequentially (non-iid), learner makes online predictions

- Performance metric: Mistake bounds

Algorithm A "Consistency"

Algorithm B “Halfing”

Now let’s get rid of “Realizability”. The setting is called “Agnostic learning”

Example: Stock forecasting

Alg C Weighted Majority

How do we fix "weighted majority"? Instead of discounting by 1/2, let's try discounting by 1-ϵ

Fact: For all $0 \leq x \leq 0.5$

$$-x - x^2 \leq \log(1-x) \leq -x$$

- Following the same analysis

Algorithm D: Randomized Weighted Majority

Analysis of RWM

From mistake bounds to loss minimization

- Loss function

- Regret

- The “Hedge” Algorithm:

Checkpoint: Online Learning

- Learning with expert advice
  - A summary of regret bound: # mistakes - Oracle # of mistakes

	Consistency	Halfing	Weighted Majority	Randomized WM
Realizable setting	$\min(T, \lvert\mathcal{H}\rvert)$	$\min(T, \log\lvert\mathcal{H}\rvert)$	$\min(T, \log\lvert\mathcal{H}\rvert)$	$\min(T, \log\lvert\mathcal{H}\rvert)$
Agnostic setting	n.a.	n.a.	$(1+\epsilon)m + \log\lvert\mathcal{H}\rvert/\epsilon$	$\sqrt{m\log\lvert\mathcal{H}\rvert} = O(\sqrt{T\log\lvert\mathcal{H}\rvert})$

Next lecture

- Online Learning (Part II)
  - Online Gradient Descent

- Reinforcement Learning
  - Markov Decision Processes