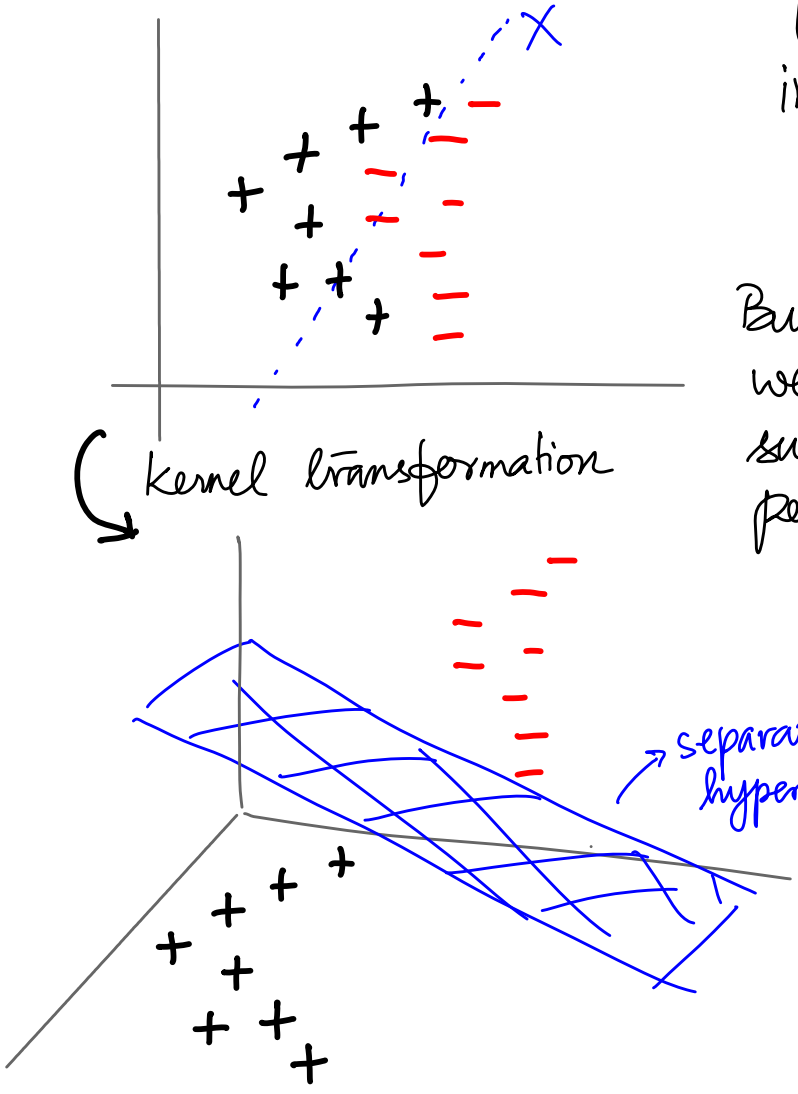
MATH 565: Lecture 22 (04/02/2026)

Today:
* Lagrangian relaxation of unconstrained problems
* Linear Regression
* Norm-constrained optimization

A Quick Note on Kernel Methods

The optimal dual expressions for H-SVM (hinge loss SVM) had terms of the form $\bar{x}_i^T \bar{x}_j$, which is the default notion/measure of similarity between data vectors $\bar{x}_i$ and $\bar{x}_j$. We could replace $\bar{x}_i^T \bar{x}_j$ with $k(\bar{x}_i, \bar{x}_j)$, a "kernel" that defines a more general measure of similarity between $\bar{x}_i$ and $\bar{x}_j$.

Use of kernels: an illustration



We cannot separate − and + instances here with a line. (perfectly, that is)

But with an appropriate kernel, we could map the points to 3D such that a plane can now perfectly separate the two classes.

The Lagrangian dual approach makes it possible to apply kernels in a straightforward manner.

We'll discuss some details in the next lecture...

Lagrangian Relaxation of Unconstrained Problems

Lagrangian relaxation and duality were naturally defined for constrained optimization — the constraints were relaxed and added as terms to the objective function with dual multipliers.

Since there are no constraints to relax in unconstrained problems, how could we go about writing its Lagrangian dual?

One approach is to add new variables as substitutions for certain terms in the objective function, and then use the constraints to define them for writing the Lagrangian dual.

We illustrate this approach on linear regression with regularization.

But first, why would we want to write the Lagrangian dual of an unconstrained problem to start with?

* To obtain lower bounds more easily;
  the dual problems are usually convex, even when the primal problems are not.
* To derive (new) optimality conditions, which could be used to develop algorithms different from ones applied to the primal problem.
* We can decouple implicit constraints, or can split variables.
  e.g., $\min J(\bar{w}) = G_1(\bar{w}) + I_C(\bar{w})$ where
  $I_C(\bar{w})$ = indicator of $\bar{w} \in C$ or not.
  → Even though $J(\cdot)$ is constrained on its own, we do need constraints to specify C. The dual could help us decouple $G_1(\bar{w})$ from these constraints.
* We can apply kernel methods.

Lagrangian Dual of Linear Regression

We illustrate the process on linear regression with regularization, which is an unconstrained minimization problem. We introduce new variables ξ_i for the error in estimating the i^{th} instance, and use its definition constraints to construct the relaxation.

$$(P) \quad \min_{\bar{w}} \; J = \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{w}^T\bar{x}_i)^2 + \frac{\lambda}{2}\|\bar{w}\|^2$$

(P) ↓ primal

With $\xi_i = y_i - \bar{w}^T\bar{x}_i$ capturing the error of fit (or prediction), we can write (P) as

$$\min \; J = \frac{1}{2}\sum \xi_i^2 + \frac{\lambda}{2}\|\bar{w}\|^2 \qquad (P)$$

$$s.t. \quad \xi_i = y_i - \bar{w}^T\bar{x}_i, \; i = 1, \ldots, n$$

D: $n \times d$ data matrix

$\bar{x}_i \in \mathbb{R}^d$ is the i^{th} data vector,

$\bar{y} \in \mathbb{R}^d$

α_i are urs (unrestricted in sign) here, since the constraints are equations

The Lagrangian relaxation can now be written as

$$\min L(\bar{w}, \bar{\xi}, \bar{\alpha}) = \frac{1}{2}\sum \xi_i^2 + \frac{\lambda}{2}\|\bar{w}\|^2 + \sum_{i=1}^{n} \alpha_i(-\xi_i + y_i - \bar{w}^T\bar{x}_i).$$

We apply stationarity conditions (to eliminate primal variables $\bar{w}, \bar{\xi}$):

$$\nabla_{\bar{w}} L = \bar{0} \text{ and } \frac{\partial L}{\partial \xi_i} = 0 \; \forall i.$$

$$\Rightarrow \lambda\bar{w} - \sum_{i=1}^{n}\alpha_i\bar{x}_i = \bar{0} \quad \text{and}$$

$$\xi_i - \alpha_i = 0 \; \forall i.$$

$$\Rightarrow \bar{w} = \frac{1}{\lambda}\sum_{i=1}^{n}\alpha_i\bar{x}_i \quad \text{and} \quad \xi_i = \alpha_i.$$

we use these equations to eliminate $\bar{w}$ and $\bar{\xi}$ from L.

$$L(\bar{w},\bar{\xi},\bar{\alpha}) = \frac{1}{2}\sum \xi_i^2 + \frac{\lambda}{2}\|\bar{w}\|^2 + \sum_{i=1}^{n} \alpha_i\left(-\xi_i + y_i - \bar{w}^T\bar{x}_i\right)$$

$$\Rightarrow L(\bar{\alpha}) = \frac{1}{2}\sum \alpha_i^2 + \frac{1}{2\lambda}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j \bar{x}_i^T\bar{x}_j + \sum_{i=1}^{n} \alpha_i\left(-\alpha_i + y_i - \frac{\bar{x}_i^T}{\lambda}\sum_{j=1}^{n}\alpha_j\bar{x}_j\right)$$

$$= \sum_{i=1}^{n}\alpha_i y_i - \frac{1}{2}\sum \alpha_i^2 - \frac{1}{2\lambda}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\bar{x}_i^T\bar{x}_j$$

$D_{n\times d} \Rightarrow DD^T$ is $n\times n$, $\bar{\alpha}\in\mathbb{R}^n$

$$\Rightarrow L(\bar{\alpha}) = \bar{\alpha}^T\bar{y} - \frac{1}{2}\|\bar{\alpha}\|^2 - \frac{1}{2\lambda}\bar{\alpha}^T DD^T\bar{\alpha}$$

$$= \bar{\alpha}^T\bar{y} - \frac{1}{2\lambda}\left(\bar{\alpha}^T\left(DD^T + \lambda I\right)\bar{\alpha}\right)$$

The dual problem can be written as $\max_{\bar{\alpha}} L(\bar{\alpha})$ (D).

First order optimality conditions give $\nabla_{\bar{\alpha}} L(\bar{\alpha}) = \bar{0}$.

$$\Rightarrow \left(DD^T + \lambda I\right)\bar{\alpha} = \lambda\bar{y}$$

$$\Rightarrow \bar{\alpha} = \lambda\left(DD^T + \lambda I\right)^{-1}\bar{y} \rightarrow \text{optimal dual solution}$$

$\Rightarrow$ We get the optimal primal solution as

$$\bar{w} = \frac{1}{\lambda}\sum_{i=1}^{n}\alpha_i\bar{x}_i = \frac{1}{\lambda}D^T\bar{\alpha}$$

$$\Rightarrow \bar{w} = D^T\left(DD^T + \lambda I\right)^{-1}\bar{y}$$

(I_n)

Can show that this expression for $\bar{w}$ is equal to the one derived in Lecture 2!

Norm Constrained Optimization

Shows up in PCA, SVD, spectral clustering, etc. Here is a typical example.

$$(P) \quad \min \sum_{i=1}^{k} \bar{x}_i^T A \bar{x}_i$$
$$\text{s.t.} \quad \|\bar{x}_i\|^2 = 1, \; i = 1, \ldots, k$$
$$\bar{x}_i \perp \bar{x}_j \quad \forall i \neq j \quad \text{(vectors are orthogonal)}$$

A is a symmetric $d \times d$ matrix.

$\bar{x}_1, \ldots, \bar{x}_k$ are variable vectors $\in \mathbb{R}^d$ ($k \leq d$ is assumed)

Orthogonality $\bar{x}_i^T \bar{x}_j = 0 \quad \forall \; 1 \leq i < j \leq k.$

We use $-\alpha_i$ as the Lagrangian multiplier for $\|\bar{x}_i\|^2 = 1$.

The Lagrangian relaxation is

$$\min_{\substack{\bar{x}_i, i=1\ldots,k \\ \bar{x}_i \perp \bar{x}_j}} L(\bar{\alpha}) = \sum_{i=1}^{k} \bar{x}_i^T A \bar{x}_i - \sum_{i=1}^{k} \alpha_i \left(\|\bar{x}_i\|^2 - 1 \right)$$

Applying stationarity gives

$$\nabla_{\bar{x}_i} L = \bar{0} \Rightarrow A\bar{x}_i = \alpha_i \bar{x}_i \quad \forall i$$

$\Rightarrow$ α_i is among the d eigenvalues of A, and $\bar{x}_i$ is the corresponding eigenvector.

$\Rightarrow$ $\bar{x}_i \perp \bar{x}_j$ are satisfied automatically, as we get orthonormal eigenvectors.

The Lagrangian function simplifies as

$$L(\bar{\alpha}) = \min_{\bar{x}_i \perp \bar{x}_j} \sum_{i=1}^{k} \alpha_i \underbrace{\bar{x}_i^T \bar{x}_i}_{=1} - \sum \alpha_i \underbrace{\left(\|\bar{x}_i\|^2 - 1\right)}_{=1}$$

$$= \min_{\substack{\text{eigenvalues} \\ \text{of } A}} \sum_{i=1}^{k} \alpha_i$$

The problem reduces to that of finding the sum of the smallest k eigenvalues of A!

Note that we assumed only that A is symmetric — we did not assume it is PSD, in particular. Hence, the objective function is not guaranteed to be convex, and hence strong duality is not guaranteed. At the same time, the optimal solution (as obtained from duality) does give the same objective function for both the primal and dual problems, thus guaranteeing optimality.