Lecture 2: Dual Support Vector Machine

1. Motivation of Dual SVM

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{s. t.} \quad y_n(\mathbf{w}^T \underbrace{\mathbf{z}_n}_{\Phi(\mathbf{x}_n)} + b) \geq 1,$$

$$\text{for } n = 1, 2, \ldots, N$$

Non-Linear Hard-Margin SVM

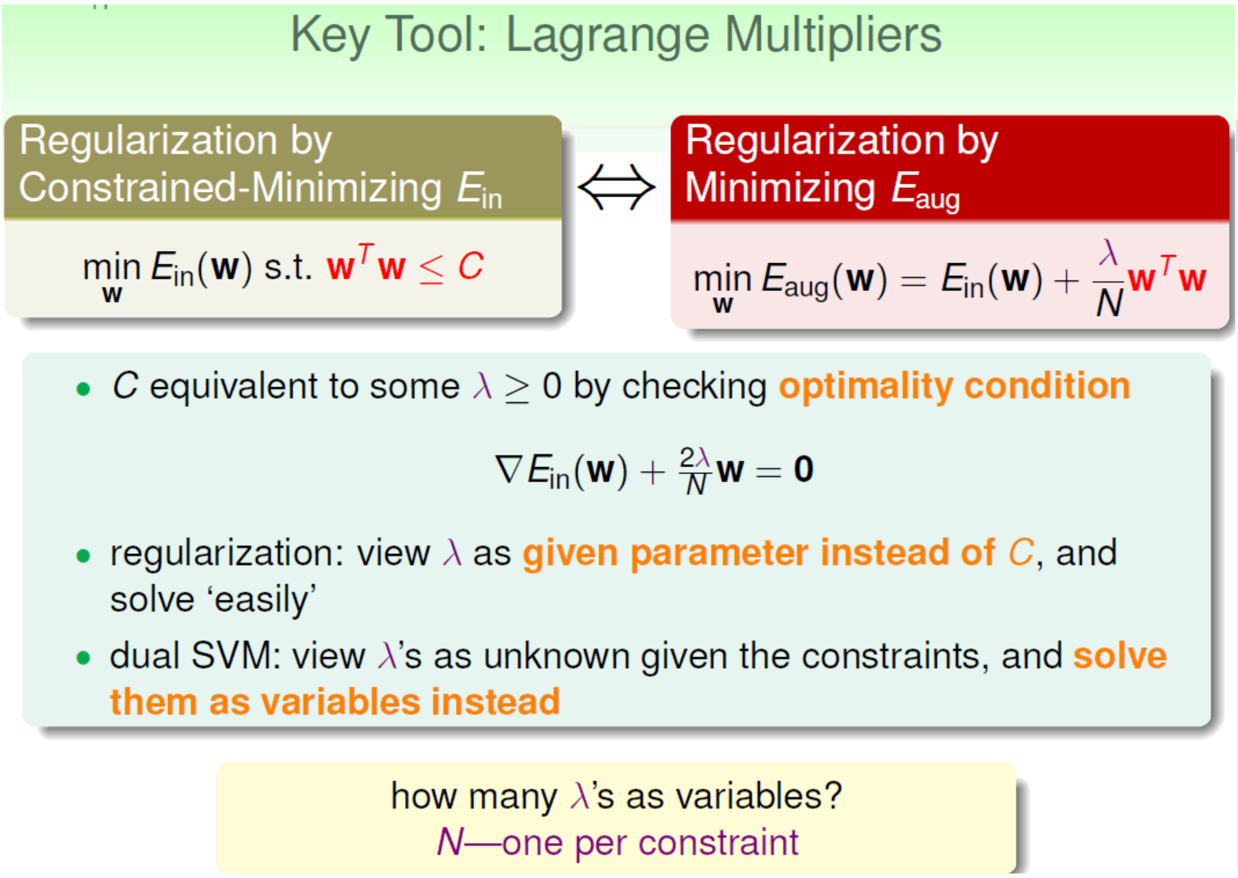
1. $\mathrm{Q} = \begin{bmatrix} 0 & \mathbf{0}_{\tilde{d}}^T \\ \mathbf{0}_{\tilde{d}} & \mathrm{I}_{\tilde{d}} \end{bmatrix}; \mathbf{p} = \mathbf{0}_{\tilde{d}+1};$ $\mathbf{a}_n^T = y_n \begin{bmatrix} 1 & \mathbf{z}_n^T \end{bmatrix}; c_n = 1$
2. $\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \leftarrow \mathrm{QP}(\mathrm{Q}, \mathbf{p}, \mathrm{A}, \mathbf{c})$
3. return $b \in \mathbb{R}$ & $\mathbf{w} \in \mathbb{R}^{\tilde{d}}$ with $g_{\mathrm{SVM}}(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^T\Phi(\mathbf{x}) + b)$

对于非线性的SVM，我们可以通过非线性的变换，将变量从x域转换到z域，在z空间中，使用线性SVM 解决问题：用二次规划的方法求出b, w,做线性分类，然后再转换为x域，就能得到最终的非线性的SVM。

使用SVM得到large-margin，减少了有效的VC Dimension，限制了模型复杂度；另一方面，使用特征转换，目的是让模型更复杂，减小E_{in}。所以说，非线性SVM是把这两者目的结合起来，平衡这两者的关系。那么，特征转换下，求解QP问题在z域中的维度设为$\hat{d}+1$，如果模型越复杂，则$\hat{d}+1$越大，相应求解这个QP问题也变得很困难。当无限大的时候，问题将会变得难以求解。

Original SVM	'Equivalent' SVM
(convex) QP of	(convex) QP of
• $\tilde{d}+1$ variables	• N variables
• N constraints	• $N+1$ constraints

Original SVM 二次规划问题的变量个数是$\hat{d}+1$ 个，N个约束条件。把该问题转换为对偶问题后，求解二次规划时，变量个数变为N个，约束条件为N+1个，与维度$\hat{d}$无关，这样就不会存在当$\hat{d}$无限大时无法求解为的情况。

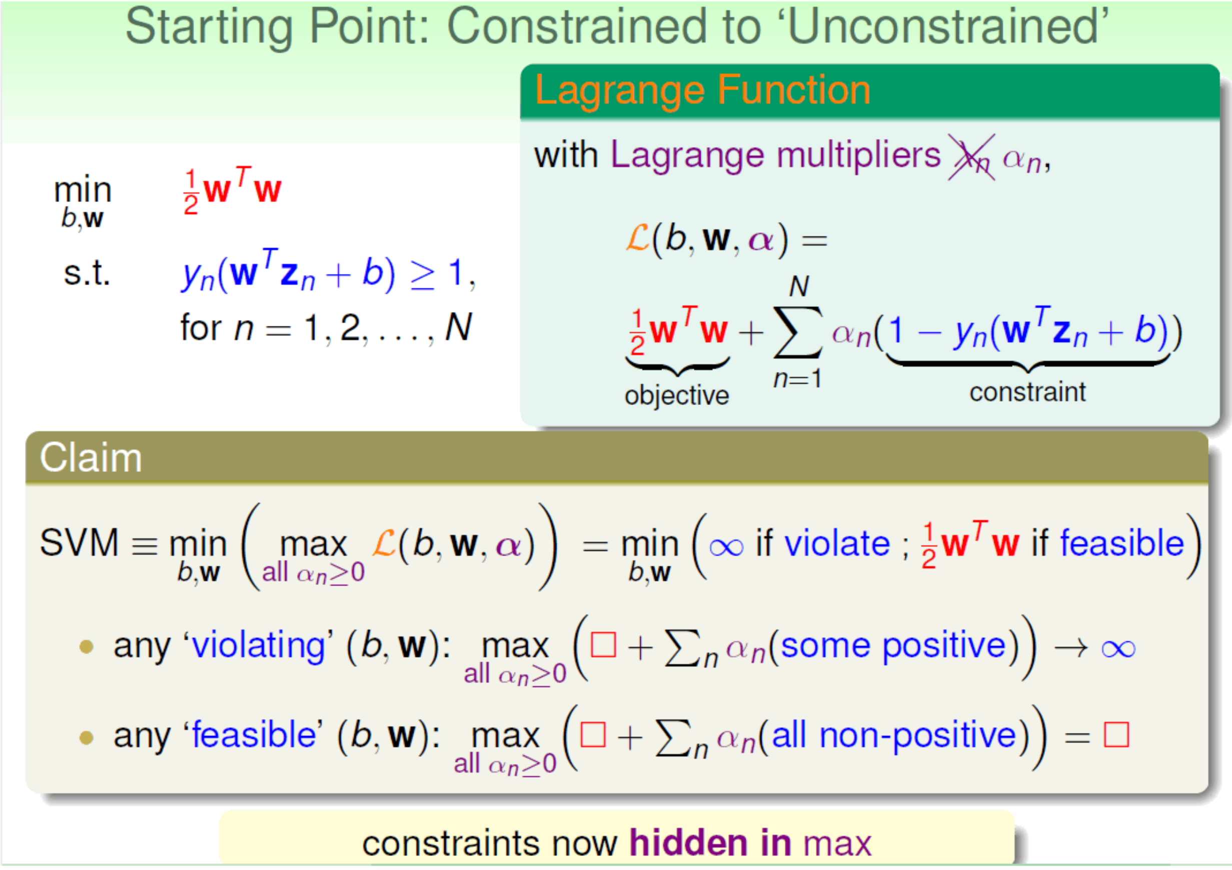


在正则化中，在最小化E_{in}的过程中，添加了限制条件$w^T w \le c$,为了将有条件的最小化问题转换为无条件的最小化问题,引入拉格朗日因子λ,得到

$$\min_{w} E_{aug}(w) = E_{in}(W) + \frac{\lambda}{N} w^T w$$

，求得最优解。

所以，在regularization问题中，λ是已知常量，求解过程变得容易。那么，对于dual SVM问题，同样可以引入λ，将条件问题转换为非条件问题，只不过λ是未知参数，且个数是N，需要对其进行求解.



首先，设定$\alpha_n \geq 0$，根据SVM的约束条件可得$1 - y_n(w^T z_n + b) \leq 0$,如果不满足最优解，即$1 - y_n(w^T z_n + b) \geq 0$,又因$\alpha_n \geq 0$,所以存在$\sum_n \alpha_n (1 - y_n(w^T z_n + b)) \geq 0$,此时最大值是趋于无穷大，是无解的。当所有的点均满足$1 - y_n(w^T z_n + b) \leq 0$，则$\sum_n \alpha_n (1 - y_n(w^T z_n + b)) \leq 0$,当$\sum_n \alpha_n (1 - y_n(w^T z_n + b)) = 0$时可得最大值$\frac{1}{2} w^T w$,这正是我们SVM的目标。因此，这种转化为非条件的SVM
构造函数的形式是可行的。

2. Lagrange Dual Problem

对于固定的α'，

$$\max_{all\ \alpha_n \geq 0} \mathcal{L}(b, w, \alpha) \geq \mathcal{L}(b, w, \alpha')$$

Lagrange Dual Problem

for any fixed α' with all $\alpha'_n \geq 0$,

$$\min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right) \geq \min_{b,\mathbf{w}} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}')$$

because max $\geq$ any

for best $\boldsymbol{\alpha}' \geq \mathbf{0}$ on RHS,

$$\min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right) \geq \underbrace{\max_{\text{all } \alpha_n' \geq 0} \min_{b,\mathbf{w}} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}')}_{\text{Lagrange dual problem}}$$

because best is one of any

Lagrange dual problem:
'outer' maximization of $\boldsymbol{\alpha}$ on lower bound of original problem

上述不等式表明，我们对SVM的min和max做了对调，满足这样的关系，这叫做Lagrange dual problem。不等式右边是SVM问题的下界，我们接下来的目的就是求出这个下界.

Strong Duality of Quadratic Programming

$$\underbrace{\min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right)}_{\text{equiv. to original (primal) SVM}} \geq \underbrace{\max_{\text{all } \alpha_n \geq 0} \left(\min_{b,\mathbf{w}} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right)}_{\text{Lagrange dual}}$$

- '$\geq$': weak duality
- '$=$': **strong duality**, true for QP if
  - convex primal
  - feasible primal (true if $\boldsymbol{\Phi}$-separable)
  - linear constraints

—called constraint qualification

exists primal-dual optimal
solution $(b, \mathbf{w}, \boldsymbol{\alpha})$ for both sides

已知≥是一种弱对偶关系，在二次规划QP问题中，如果满足以下三个条件：

- 函数是凸的（convex primal）
- 函数有解（feasible primal）
- 条件是线性的（linear constraints）

那么，上述不等式关系就变成强对偶关系，$\geq$变成$=$，即一定存在满足条件的解(b, w, α)，使等式左边和右边都成立，SVM的解就转化为右边的形式。
经过推导，SVM对偶问题的解已经转化为无条件形式：

$$\max_{\text{all } \alpha_n \geq 0} \left(\underbrace{\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)\right)}_{\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})} \right)$$

- inner problem 'unconstrained', at optimal:
  $\frac{\partial \mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})}{\partial b} = 0 = -\sum_{n=1}^{N} \alpha_n y_n$
- no loss of optimality if solving with constraint $\sum_{n=1}^{N} \alpha_n y_n = 0$

其中，上式括号里面的是对拉格朗日函数$\mathcal{L}(b, w, \alpha)$计算最小值。那么根据梯度下降算法思想：最小值位置满足梯度为零。首先，令$\mathcal{L}(b, w, \alpha)$对参数b的梯度为零,得到$\sum_{n=1}^{N} \alpha_n y_n = 0$，带入原式化简为：

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0} \left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y_n(\mathbf{w}^T\mathbf{z}_n)\right) \right)$$

然后，再根据最小值思想，令$\mathcal{L}(b, w, \alpha)$对参数$w$的梯度为零,得到：

- inner problem 'unconstrained', at optimal:
  $\frac{\partial \mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})}{\partial w_i} = 0 = w_i - \sum_{n=1}^{N} \alpha_n y_n z_{n,i}$
- no loss of optimality if solving with constraint $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n$

将其带入原式化简为：

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} \left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n - \mathbf{w}^T\mathbf{w} \right)$$

$$\Longleftrightarrow \max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\left\| \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n \right\|^2 + \sum_{n=1}^{N} \alpha_n$$

这样，SVM表达式消去了w，SVM最佳化形式转化为只与α有关。

KKT Optimality Conditions

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\|\sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n\|^2 + \sum_{n=1}^{N} \alpha_n$$

if primal-dual optimal $(b, \mathbf{w}, \boldsymbol{\alpha})$,

- primal feasible: $y_n(\mathbf{w}^T \mathbf{z}_n + b) \geq 1$
- dual feasible: $\alpha_n \geq 0$
- dual-inner optimal: $\sum y_n \alpha_n = 0$; $\mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n$
- primal-inner optimal (at optimal all 'Lagrange terms' disappear):

$$\alpha_n(1 - y_n(\mathbf{w}^T \mathbf{z}_n + b)) = 0$$

—called **Karush-Kuhn-Tucker (KKT) conditions**, necessary for optimality [& sufficient here]

will use **KKT** to 'solve' $(b, \mathbf{w})$ from optimal $\boldsymbol{\alpha}$

原问题的约束
对偶问题的约束
对偶问题内部求最佳化的约束

3. Solving Dual SVM

Dual Formulation of Support Vector Machine

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\|\sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n\|^2 + \sum_{n=1}^{N} \alpha_n$$

standard hard-margin SVM dual

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - \sum_{n=1}^{N} \alpha_n$$

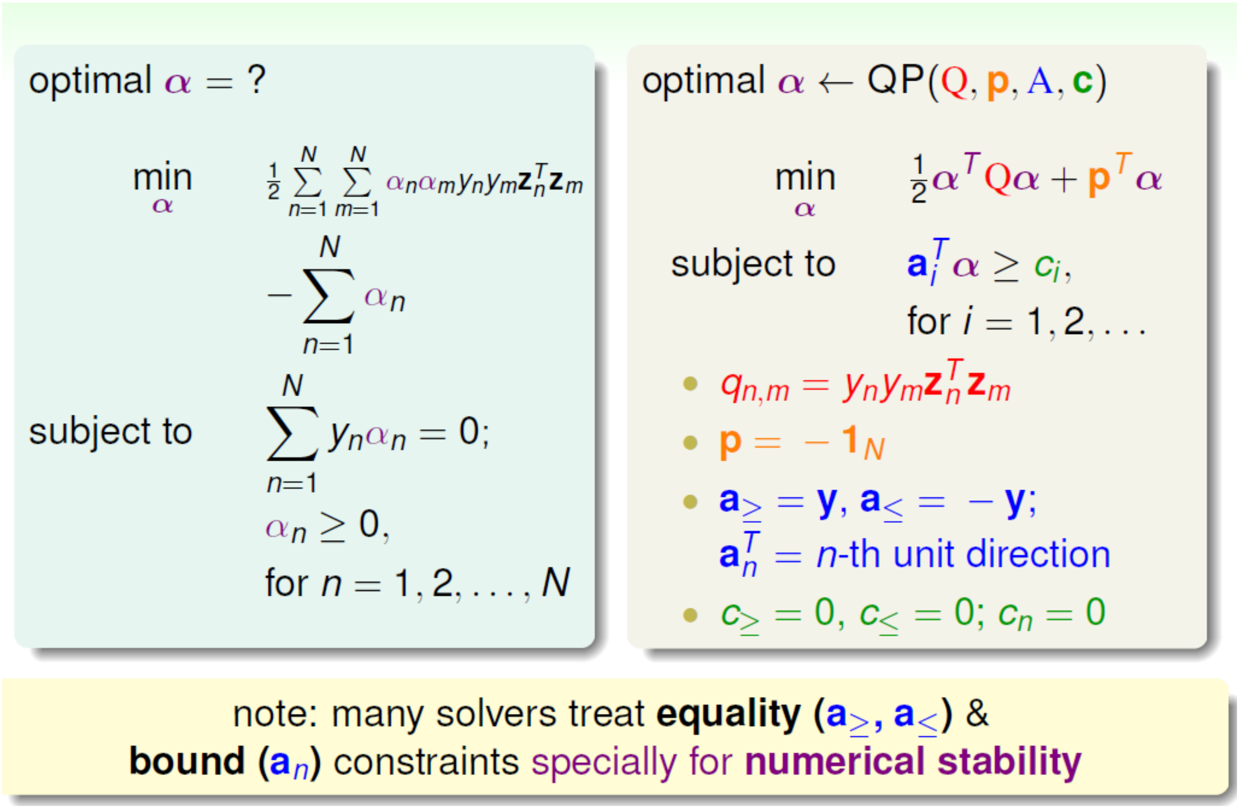
$$\text{subject to} \quad \sum_{n=1}^{N} y_n \alpha_n = 0;$$

$$\alpha_n \geq 0, \text{ for } n = 1, 2, \ldots, N$$

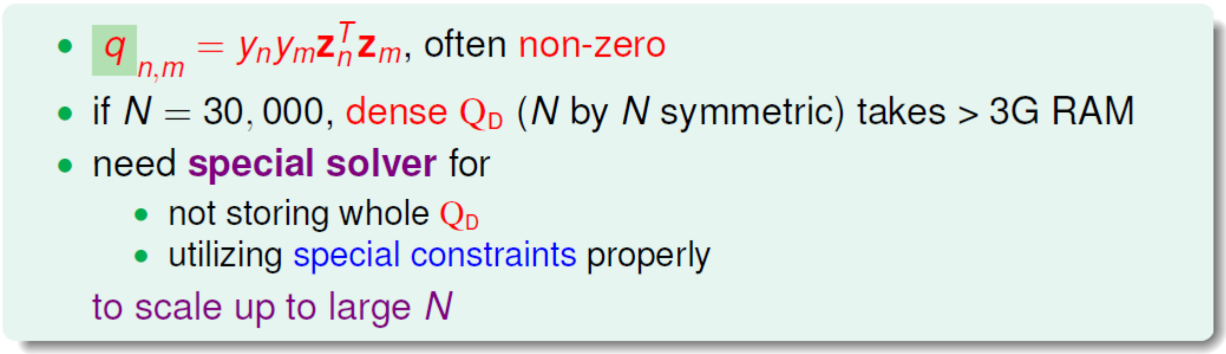
(convex) QP of N variables & $N + 1$ constraints, as promised

二次规划里面的系数含义：

Q：二次项系数 p：一次项系数 A：条件里面的系数 c:条件里面的常数



在求解过程中,$q_{n,m}=y_n y_m z_n^T z_m$大部分值是非零的，称为dense。当N很大的时候，例如N=30000，那么对应的的计算量将会很大，存储空间也很大。所以一般情况下，对dual SVM问题的矩阵，需要使用一些特殊的方法.



 通过计算得到α后，根据KKT条件，可以计算出b,w

通过$w=\sum\alpha_n y_n z_n$可以得到w,然后利用$\alpha_n(1-y_n(w^T z_n+b))=0$约束，令$\alpha_n\geq 0$,则$(1-y_n(w^T z_n+b))=0$,推出$b=y_n-w^T z_n$

Optimal $(b, \mathbf{w})$

KKT conditions

if primal-dual optimal $(b, \mathbf{w}, \alpha)$,

- primal feasible: $y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1$
- dual feasible: $\alpha_n \geq 0$
- dual-inner optimal: $\sum y_n\alpha_n = 0$; $\mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n$
- primal-inner optimal (at optimal all 'Lagrange terms' disappear):

$$\alpha_n(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)) = 0 \text{ (complementary slackness)}$$

- optimal $\alpha \Longrightarrow$ optimal $\mathbf{w}$? easy above!
- optimal $\alpha \Longrightarrow$ optimal b? a range from primal feasible & equality from **comp. slackness** if one $\alpha_n > 0 \Rightarrow b = y_n - \mathbf{w}^T\mathbf{z}_n$

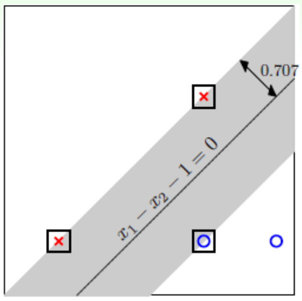
comp. slackness:
$\alpha_n > 0 \Rightarrow$ **on fat boundary (SV!)**

在计算b值的时候，$\alpha_n \geq 0$时,有$y_n(w^T z_n + b)) = 1$,表示该点在SVM的分类线上，是fat boundary。

4. Messages behind Dual SVM

一开始我们把在边界上的点称为支撑向量，在解决了对偶问题后，如果$\alpha_n \geq 0$，那它一定在边界上，称为support vectors,candidates 称为在边界上的点。

- on boundary: 'locates' fattest hyperplane; others: **not needed**
- examples with $\alpha_n > 0$: on boundary
- call $\alpha_n > 0$ examples $(\mathbf{z}_n, y_n)$ **support vectors** ~~(candidates)~~
- SV (positive α_n) $\subseteq$ SV candidates (on boundary)



- only SV needed to compute $\mathbf{w}$: $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n = \sum_{\text{SV}} \alpha_n y_n \mathbf{z}_n$
- only SV needed to compute b: $b = y_n - \mathbf{w}^T\mathbf{z}_n$ with any SV $(\mathbf{z}_n, y_n)$

SV只由$\alpha_n > 0$的点决定，根据上一部分推导的w和b的计算公式，我们发现，w和b仅由SV即$\alpha > 0$的点决定，简化了计算量。这跟我们上一节课介绍的分类线只由“胖”边界上的点所决定是一个道理。也就是说，样本点可以分成两类：一类是support vectors，通过support vectors可以求得fattest hyperplane；另一类不是support vectors，对我们求得fattest hyperplane没有影响。

Representation of Fattest Hyperplane

SVM

$$\mathbf{w}_{\text{SVM}} = \sum_{n=1}^{N} \alpha_n (y_n \mathbf{z}_n)$$

α_n from **dual solution**

PLA

$$\mathbf{w}_{\text{PLA}} = \sum_{n=1}^{N} \beta_n (y_n \mathbf{z}_n)$$

β_n by **# mistake corrections**

$\mathbf{w}$ = linear combination of $y_n \mathbf{z}_n$

- also true for GD/SGD-based LogReg/LinReg when $\mathbf{w}_0 = \mathbf{0}$
- call $\mathbf{w}$ **'represented' by data**

SVM: represent $\mathbf{w}$ by **SVs only**

我们发现，二者在形式上是相似的。w_{SVM}由fattest hyperplane边界上所有的SV决定，w_{PLA}由所有当前分类错误的点决定。w_{SVM}和w_{PLA}都是原始数据点的线性组合形式，是原始数据的代表。

Summary: Two Forms of Hard-Margin SVM

Primal Hard-Margin SVM

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{sub. to} \quad y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1, \quad \text{for } n = 1, 2, \ldots, N$$

- $\tilde{d}+1$ variables, N constraints —suitable when $\tilde{d}+1$ small
- physical meaning: locate **specially-scaled** $(b, \mathbf{w})$

Dual Hard-Margin SVM

$$\min_{\alpha} \quad \frac{1}{2}\alpha^T \mathrm{Q}_{\mathrm{D}} \alpha - \mathbf{1}^T \alpha$$

$$\text{s.t.} \quad \mathbf{y}^T\alpha = 0; \quad \alpha_n \geq 0 \text{ for } n = 1, \ldots, N$$

- N variables, $N+1$ simple constraints —suitable when N small
- physical meaning: locate **SVs** $(\mathbf{z}_n, y_n)$ & their α_n

both eventually result in optimal $(b, \mathbf{w})$ for fattest hyperplane

$$g_{\text{SVM}}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\boldsymbol{\Phi}(\mathbf{x}) + b)$$

总结一下，本节课和上节课主要介绍了两种形式的SVM，一种是Primal HardMarginSVM，另一种是Dual Hard_Margin SVM。Primal HardMargin
SVM有个$\hat{d}+1$参数，有N个限制条件。当$\hat{d}+1$很大时，求解困难。而Dual Hard_Margin SVM有

N个参数，有N+1个限制条件。当数据量N很大时，也同样会增大计算难度。两种形式都能得到w和b，求得fattest hyperplane。通常情况下，如果N不是很大，一般使用Dual SVM来解决问题。

$$\min_{\alpha} \quad \frac{1}{2}\alpha^T \mathrm{Q}_{\mathrm{D}} \alpha - \mathbf{1}^T \alpha$$
$$\text{subject to} \quad \mathbf{y}^T \alpha = 0;$$
$$\alpha_n \geq 0, \text{for } n = 1, 2, \ldots, N$$

- N variables, $N+1$ constraints: **no dependence on $\tilde{d}$?**
- $q_{n,m} = y_n y_m \mathbf{z}_n^T \mathbf{z}_m$: inner product in $\mathbb{R}^{\tilde{d}}$
  —$O(\tilde{d})$ via naïve computation!

Dual SVM是否真的消除了对$\hat{d}$的依赖呢？其实并没有。因为在计算$q_{n,m} = y_n y_m z_n z_m$的过程中，由z向量引入了$\hat{d}$，实际上复杂度已经隐藏在计算过程中了。所以，我们的目标并没有实现。

5. 总结

本节课主要介绍了SVM的另一种形式：Dual SVM。我们这样做的出发点是为了移除计算过程对$\hat{d}$的依赖。Dual SVM的推导过程是通过引入拉格朗日因子α，将SVM转化为新的非条件形式。然后，利用QP，得到最佳解的拉格朗日因子α。再通过KKT条件，计算得到对应的w和b。最终求得fattest hyperplane。