Project Presentations

Tuesday, 3/1

1. Group 1 (**Hussain**, Tanvir; **Lewis**, Cameron; **Villamar**, Sandra)
2. Group 2 (**Dong**, Meng; **Long**, Jianzhi; **Wen**, Bo; **Zhang**, Haochen)
3. Group 3 (**Chen**, Yuzhao; **Li**, Zonghuan; **Song**, Yuze; **Yan**, Ge)
4. Group 4 (**Li**, Jiayuan; **Xiao**, Nan; **Yu**, Nancy; **Zhou**, Pei)
5. Group 5 (**Li**, Zheng; **Tao**, Jianyu; **Yang**, Fengqi)
6. Group 6 (**Bian**, Xintong; **Jiang**, Yufan; **Wu**, Qiyao)
7. Group 7 (**Chen**, Yongxing; **Yao**, Yanzhi; **Zhang**, Canwei)
8. Group 8 (**Nukala**, Kishore; **Pulleti**, Sai; **Vaidyula**, Srikar)

Thursday, 3/3

1. Group 9 (**Baluja**, Michael; **Cao**, Fangning; **Huff**, Mikael; **Shen**, Xuyang)
2. Group 10 (**Arun**, Aditya; **Long**, Heyang; **Peng**, Haonan)
3. Group 11 (**Cowin**, Samuel; **Liao**, Albert; **Mandadi**, Sumega)
4. Group 12 (**Jia**, Yichen; **Jiang**, Zhiyun; **Li**, Zhuofan)
5. Group 13 (**Dandu**, Murali; **Daru**, Srinivas; **Pamidi**, Sri)
6. Group 14 (**He**, Bolin; **Huang**, Yen-Ting; **Wang**, Shi; **Wang**, Tzu-Kao)
7. Group 15 (**Chen**, Luobin; **Feng**, Ruining; **Wu**, Ximei; **Xu**, Haoran)

Tuesday, 3/8

1. Group 16 (**Chen**, Rex; **Liang**, Youwei; **Zheng**, Xinran)
2. Group 17 (**Aguilar**, Matthew; **Millhiser**, Jacob; **O'Boyle**, John; **Sharpless**, Will)
3. Group 18 (**Wang**, Haoyu; **Wang**, Jiawei; **Zhang**, Yuwei)
4. Group 19 (**Chen**, Yinbo; **Di**, Zonglin; **Mu**, Jiteng)
5. Group 20 (**Chowdhury**, Debalina; **He**, Scott; **Ye**, Yiheng)
6. Group 21 (**Lin**, Wei-Ru; **Ru**, Liyang; **Zhang**, Shaohua)
7. Group 22 (**Bhavsar**, Shivad; **Blazej**, Christopher; **Bu**, Yinyan; **Liu**, Haozhe)

Thursday, 3/10

1. Group 23 (**Chen**, Claire; **Hsieh**, Chia-Wei; **Lin**, Jui-Yu; **Tsai**, Ya-Chen)
2. Group 24 (**Cheng**, Yu; **Yu**, Zhaowei; **Zaidi**, Ali)
3. Group 25 (**Assadi**, Parsa; **Brugere**, Tristan; **Pathak**, Nikhil; **Zou**, Yuxin)
4. Group 28 (**Candassamy**, Gokulakrishnan; **Dixit**, Rajeev; **Huang**, Joyce)
5. Group 27 (**Kok**, Hong; **Wang**, Jacky; **Yan**, Yijia; **Yuan**, Zhouyuan)
6. Group 28 (**Luan**, Zeting; **Yang**, Zheng)
7. Group 29 (**Cuawenberghs**, Kalyani; **Mojtahed**, Hamed)

Each presentation will be allocated 9 minutes (<u>pts will be deducted if you go over 9 minutes</u>)

The presentation slides of <u>ALL GROUPS</u> are due by Monday, 2/28 @ 11:59 pm

Email me the file (mvasconcelos@eng.ucsd.edu) and **<u>name the file</u>** GroupX.pdf, where X is your group number (see previous slide). Use Group X Presentation as the **<u>subject of your email</u>** and **<u>cc to all members</u>**.

The presentation should discuss the **problem that you are trying to solve**, the **data that you are using**, the **proposed solution**(s), and the **results that you have so far** (<u>they can later be UPDATED IN THE PROJECT PAPER</u>).

ECE 271B – Winter 2022

The Soft–Margin Support Vector Machine

Disclaimer:
This class will be recorded
and made available to students asynchronously.

Manuela Vasconcelos

ECE Department, UCSD

The Support Vector Machine

- the SVM is the classifier that maximizes the margin under the constraints

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \forall i$$

- no dual gap, and the dual problem is

$$\max_{\alpha \geq 0} \left\{ -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_i \alpha_i \right\} \quad \text{subject to} \quad \sum_i \alpha_i y_i = 0$$

  once this is solved, the vector

$$\mathbf{w}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i$$

  is the normal to the maximum margin plane

- **note:** the dual solution does not determine the optimal b^*

Support Vectors

- from the **KKT conditions**, a active (inactive) constraint has non−zero (zero) Lagrange multiplier α_i
- that is

$$\alpha_i > 0 \quad \text{iff} \quad y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) = 1$$

- hence $\boxed{\alpha_i > 0}$ only for points

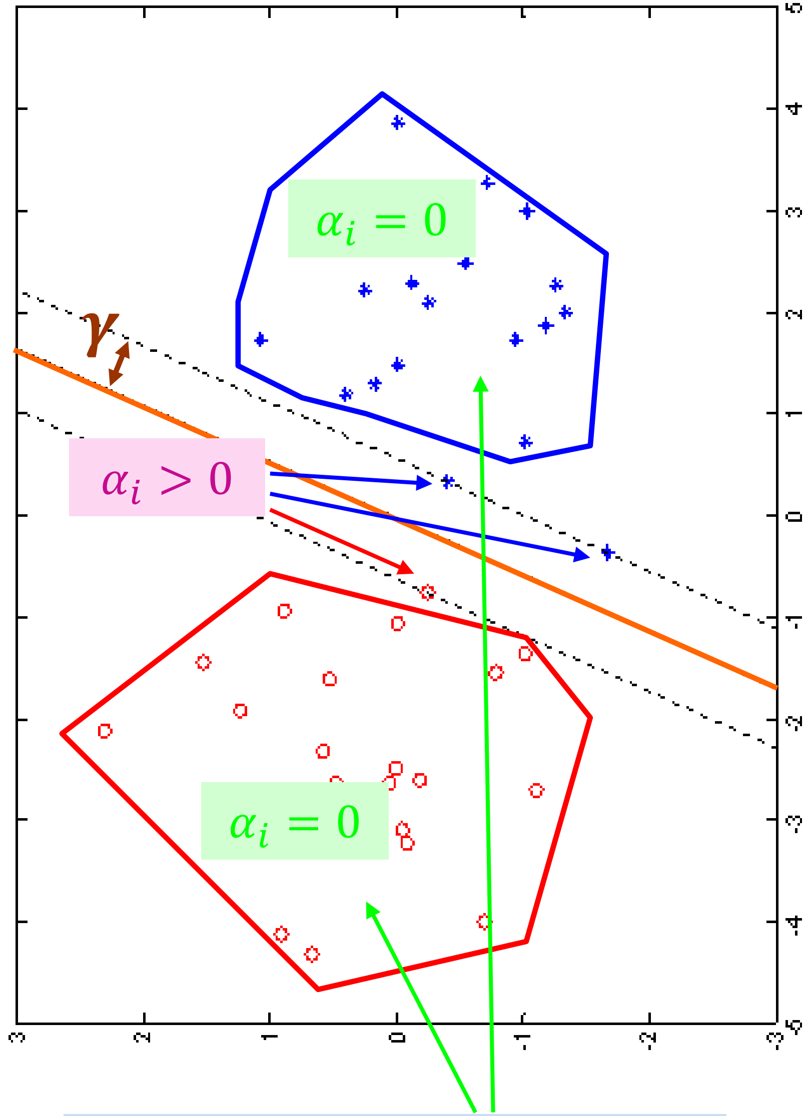
$$\boxed{\left|\mathbf{w}^{*T}\mathbf{x}_i + b^*\right| = 1}$$

which are those **that lie at a distance equal to the margin**

- these "support" the hyperplane and are called support vectors
- the **decision rule** is

$$\boxed{f(\mathbf{x}) = \operatorname{sgn}\left[\sum_{i \in SV} y_i \alpha_i^* \mathbf{x}_i^T \mathbf{x} + b^*\right]}$$

$$\boxed{SV = \{i \mid \alpha_i^* > 0\}}$$



and the remaining points are **irrelevant**!

Hard–Margin SVM

- SVM training

  1) solve the **optimization problem**

$$\boxed{\max_{\boldsymbol{\alpha}\geq 0}\left\{-\frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum_i \alpha_i\right\} \quad \text{subject to} \quad \sum_i \alpha_i y_i = 0}$$

  2) then **compute**

$$\boxed{\mathbf{w}^* = \sum_{i\in SV}\alpha_i^* y_i \mathbf{x}_i} \qquad \boxed{b^* = -\frac{1}{2}\sum_{i\in SV} y_i\alpha_i^*\left(\mathbf{x}_i^T\mathbf{x}^+ + \mathbf{x}_i^T\mathbf{x}^-\right)}$$

- decision function

$$\boxed{f(\mathbf{x}) = \operatorname{sgn}\left[\sum_{i\in SV} y_i\alpha_i^*\mathbf{x}_i^T\mathbf{x} + b^*\right]}$$

SVM: Kernelization

$$\max_{\alpha \geq 0} \left\{ -\tfrac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_i \alpha_i \right\} \quad \text{subject to} \sum_i \alpha_i y_i = 0$$

$$b^* = -\tfrac{1}{2} \sum_{i \in SV} y_i \alpha_i^* \left(\mathbf{x}_i^T \mathbf{x}^+ + \mathbf{x}_i^T \mathbf{x}^- \right)$$

$$f(\mathbf{x}) = \operatorname{sgn}\left[\sum_{i \in SV} y_i \alpha_i^* \mathbf{x}_i^T \mathbf{x} + b^* \right]$$

- note that <u>all</u> equations <u>depend **only**</u> on $\boxed{\mathbf{x}_i^T \mathbf{x}_j}$
- the "kernel trick" is **trivial**: replace by $K(\mathbf{x}_i, \mathbf{x}_j)$

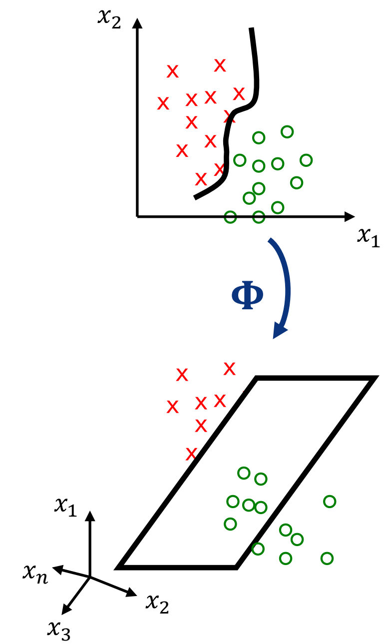
1) training

$$\boxed{\max_{\alpha \geq 0} \left\{ -\tfrac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \sum_i \alpha_i \right\} \text{ subject to} \sum_i \alpha_i y_i = 0}$$

$$\boxed{b^* = -\tfrac{1}{2} \sum_{i \in SV} y_i \alpha_i^* \left(K(\mathbf{x}_i, \mathbf{x}^+) + K(\mathbf{x}_i, \mathbf{x}^-) \right)}$$

2) decision function

$$\boxed{f(\mathbf{x}) = \operatorname{sgn}\left[\sum_{i \in SV} y_i \alpha_i^* K(\mathbf{x}_i, \mathbf{x}) + b^* \right]}$$



- note that we can <u>no longer</u> recover $\mathbf{w}^*$ <u>explicitly</u> without determining the feature transformation $\mathbf{\Phi}$, but, **luckily**, we **do <u>not</u> really need** $\mathbf{w}^*$, <u>only</u> the decision function

$$\mathbf{w}^* = \sum_{i \in SV} \alpha_i^* y_i \, \mathbf{\Phi}(\mathbf{x}_i)$$

Input–Space Interpretation

$$f(\mathbf{x}) = \operatorname{sgn}\left[\sum_{i \in SV} y_i \alpha_i^* K(\mathbf{x}_i, \mathbf{x}) + b^*\right]$$

- last class, we saw that the decision function identical to the BDR for

1) class 1 with likelihood and prior

$$\boxed{\sum_{i \in SV | y_i \geq 0} \pi_i^* K(\mathbf{x}_i, \mathbf{x})} \qquad \pi_i^* = \frac{\alpha_i^*}{\sum_{i \in SV | y_i \geq 0} \alpha_i^*}, i | y_i \geq 0 \qquad \boxed{\sum_{i \in SV | y_i \geq 0} \alpha_i^* \Big/ \sum_i \alpha_i^*}$$

2) class 2 with likelihood and prior

$$\boxed{\sum_{i \in SV | y_i < 0} \beta_i^* K(\mathbf{x}_i, \mathbf{x})} \qquad \beta_i^* = \frac{\alpha_i^*}{\sum_{i \in SV | y_i < 0} \alpha_i^*}, i | y_i < 0 \qquad \boxed{\sum_{i \in SV | y_i < 0} \alpha_i^* \Big/ \sum_i \alpha_i^*}$$

- i.e.

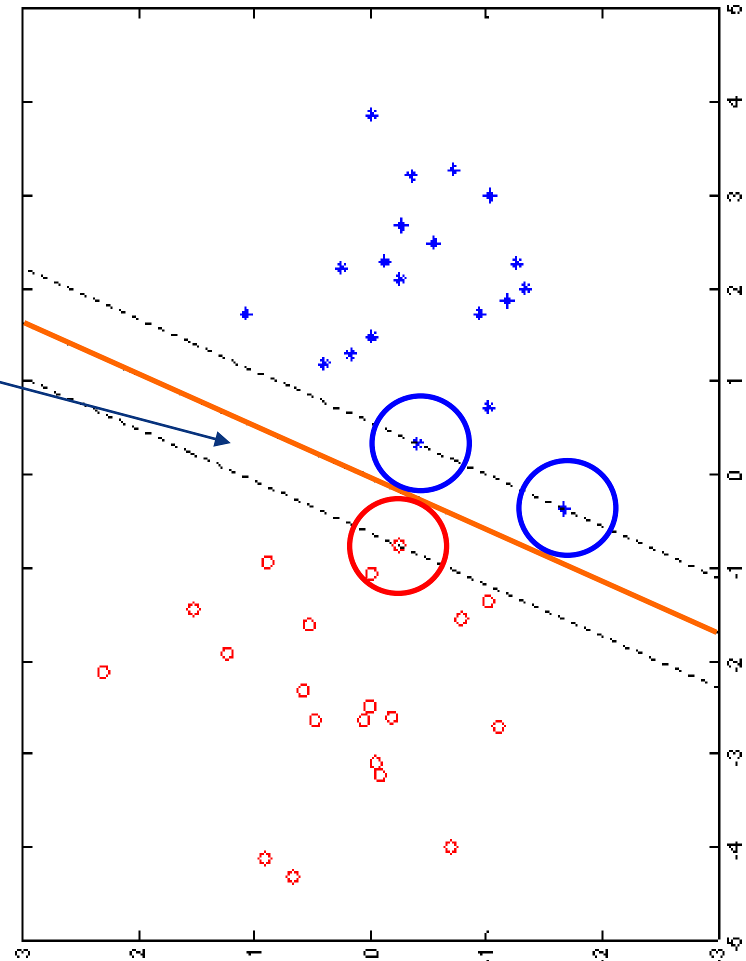
$$f(\mathbf{x}) = \begin{cases} 1, & \dfrac{\sum_{i \in SV | y_i \geq 0} \pi_i^* K(\mathbf{x}_i, \mathbf{x})}{\sum_{i \in SV | y_i < 0} \beta_i^* K(\mathbf{x}_i, \mathbf{x})} \geq T \\ -1, & \text{otherwise} \end{cases}$$

BDR threshold

- these likelihood functions are a <u>kernel</u> density estimate if $K(\cdot, \mathbf{x}_i)$ is a valid pdf

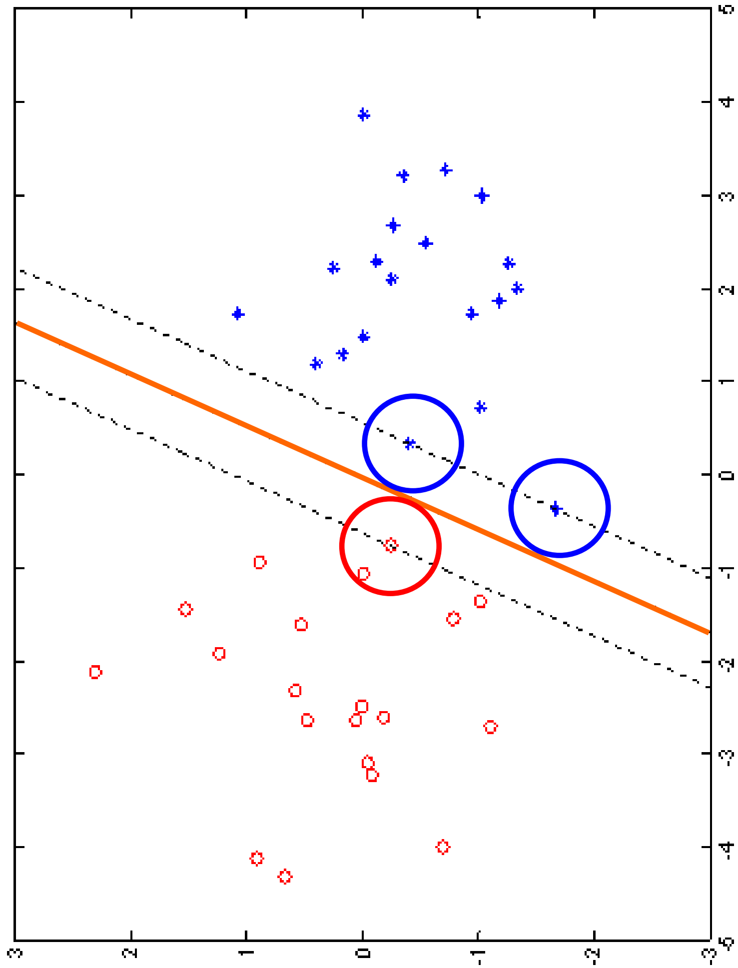
Input–Space Interpretation

- peculiar kernel estimates
  - only place kernels on the support vectors, all other points ignored
- discriminant density estimation
  - concentrate modeling power where it **matters the most**, i.e. near **classification boundary**
  - **smart**, since points away from the boundary are always well classified even if the density estimates in their region are poor
  - the **SVM** is a highly efficient combination of the **BDR** with **kernel estimates**, complexity $O(|SV|)$ instead of $O(n)$



Limitations of the SVM

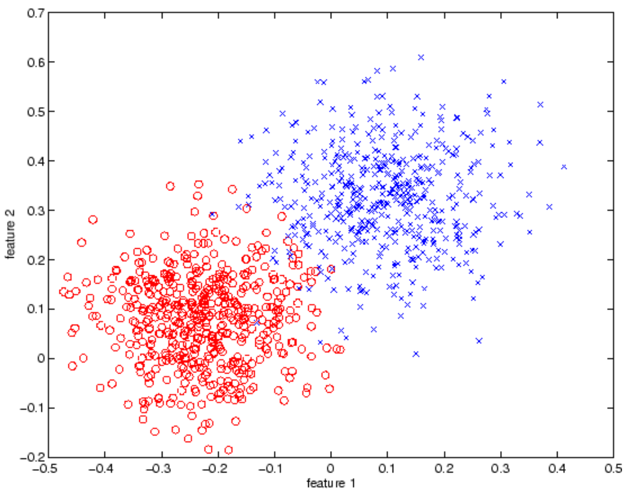
- appealing, but also points out the <u>limitations</u> of the SVM:
  - **major problem** of kernel density estimation is the choice of bandwidth
    - if too small, the estimates have too much variance
    - if too large, the estimates have too much bias
  - this problem **appears <u>again</u>** in the SVM
    - <u>no</u> generic "optimal" procedure to find the kernel or its parameters
    - requires <u>**trial and error**</u>
    - note, however, that this is **less of a headache** since only a <u>few</u> kernels have to be evaluated



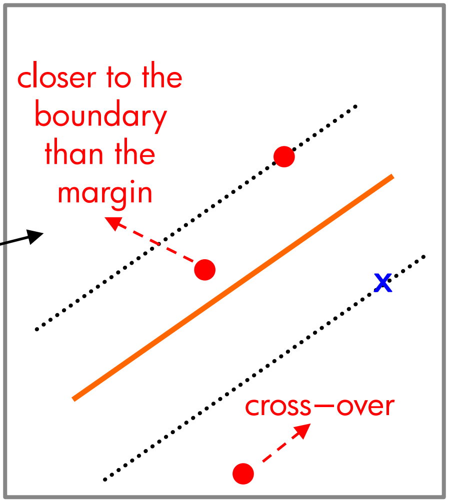
- usually, we pick an **arbitrary kernel**, e.g. Gaussian
- then, determine **kernel parameters**, e.g. variance, by <u>**trial and error**</u>

SVM: Non–Separable Problems

- so far, we have **assumed** linearly separable classes
- this is **rarely the case in practice**
- a separable problem is "**easy**": most classifiers will do well
- we need to be able to **extend** the SVM to the non–separable case



- basic idea:
  - with class overlap, we **cannot** enforce a margin
  - but we **can** enforce a soft–margin
    - for **most points**, there is a margin
    - but then there are a few outliers that cross–over or are closer to the boundary than the margin



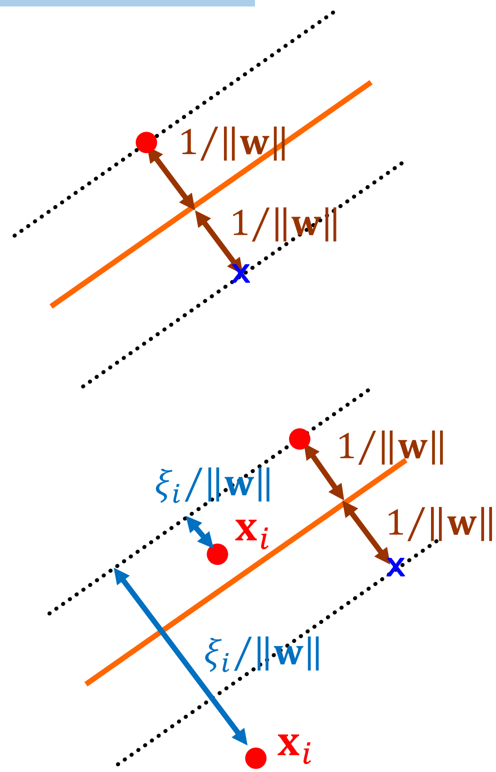
SVM: Soft–Margin Optimization

- mathematically, this can be done by introducing slack variables
- instead of solving the hard–margin problem

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \forall i$$

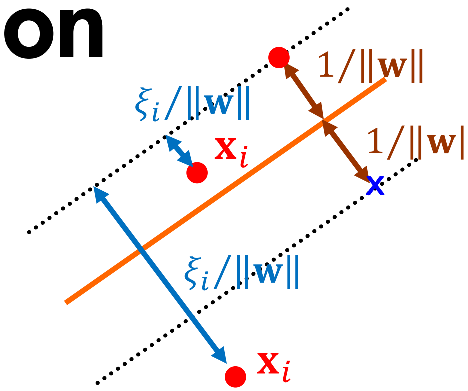
- we solve the soft–margin problem

$$\min_{\mathbf{w},\xi,b} \|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i$$
$$\xi_i \geq 0, \forall i$$



- the ξ_i are called slacks
- basically, the **same as before**, but points with $\xi_i > 0$ are allowed to violate the margin

SVM: Soft−Margin Optimization



- note that the problem is <u>not</u> really **well defined**
  - by making ξ_i arbitrarily large, any $\mathbf{w}$ will do
  - we need to <u>penalize</u> large ξ_i

- this is done by solving instead the <u>regularized optimization</u> problem

$$\min_{\mathbf{w},\xi,b} \|\mathbf{w}\|^2 + Cf(\xi)$$

$$\text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i$$

$$\xi_i \geq 0, \forall i$$

$Cf(\xi)$ – <u>penalty</u> or <u>regularization</u> term

$C > 0$ controls how <u>harsh</u> the penalty is

- $f(\xi)$ is **usually** a **norm**: we consider

the 1−**norm**

$$f(\xi) = \sum_i \xi_i$$

the 2−**norm**

$$f(\xi) = \sum_i \xi_i^2$$

2–Norm SVM

$$\min_{\mathbf{w},\boldsymbol{\xi},b} \ \|\mathbf{w}\|^2 + C\sum_i \xi_i^2$$

$$\text{subject to} \ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i \qquad (**)$$

$$\xi_i \geq 0, \forall i$$

- **note** that
  - if $\xi_i < 0$ and the constraint $(**)$ is satisfied,
    then $(**)$ is satisfied by $\xi_i = 0$ and the cost will be smaller
  - hence $\xi_i < 0$ is **never** a solution and the positivity constraints on the ξ_i are redundant
  - they can therefore be dropped

2–Norm SVM

- this leads to

$$\min_{\mathbf{w},\xi,b} \tfrac{1}{2}\|\mathbf{w}\|^2 + \tfrac{C}{2}\sum_i \xi_i^2$$
$$\text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i$$

- and

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \tfrac{1}{2}\|\mathbf{w}\|^2 + \tfrac{1}{2}C\sum_i \xi_i^2 + \sum_i \alpha_i[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)]$$

- from which

$$\nabla_{\mathbf{w}} L = 0 \quad \Leftrightarrow \quad \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0 \quad \Leftrightarrow \quad \boxed{\mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i}$$

$$\nabla_b L = 0 \quad \Leftrightarrow \quad \boxed{\sum_i \alpha_i y_i = 0} \qquad\qquad \nabla_{\xi_i} L = 0 \quad \Leftrightarrow \quad \boxed{C\xi_i - \alpha_i = 0}$$

2−Norm SVM

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \tfrac{1}{2}\|\mathbf{w}\|^2 + \tfrac{1}{2}C\sum_i \xi_i^2 + \sum_i \alpha_i[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)]$$

- plugging back

$$\mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i \qquad \sum_i \alpha_i y_i = 0 \qquad \xi_i = \frac{\alpha_i}{C}$$

- we get the Lagrangian

$$L(\mathbf{w}^*, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \tfrac{1}{2}\|\mathbf{w}^*\|^2 + \tfrac{C}{2}\sum_i \left(\tfrac{\alpha_i}{C}\right)^2 + \sum_i \alpha_i \left[1 - \frac{\alpha_i}{C} - y_i\left(\mathbf{w}^{*T}\mathbf{x}_i + b\right)\right]$$

$$= \tfrac{1}{2}\sum_{ij} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \tfrac{1}{2}\sum_i \frac{\alpha_i^2}{C} + \sum_i \alpha_i - \sum_{ij} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \underbrace{\sum_i \alpha_i y_i b}_{0}$$

$$= -\tfrac{1}{2}\left(\sum_{ij} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \frac{1}{C}\sum_i \alpha_i^2\right) + \sum_i \alpha_i$$

$$= -\tfrac{1}{2}\sum_{ij} \alpha_i\alpha_j y_i y_j \left(\mathbf{x}_i^T\mathbf{x}_j + \frac{\delta_{ij}}{C}\right) + \sum_i \alpha_i$$

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

Soft Dual for 2−Norm

hard−margin

$$\max_{\alpha \geq 0} \left\{ -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_i \alpha_i \right\}$$

$$\text{subject to} \sum_i \alpha_i y_i = 0$$

- the dual problem is

$$\boxed{\begin{gathered} \max_{\alpha \geq 0} \left\{ -\frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \left(\mathbf{x}_i^T \mathbf{x}_j + \frac{\delta_{ij}}{C} \right) + \sum_i \alpha_i \right\} \\ \text{subject to} \sum_i \alpha_i y_i = 0 \end{gathered}}$$

- **same** as hard−margin, with $\frac{1}{C}\mathbf{I}$ added to kernel matrix

$$\sum_{ij} \alpha_i \alpha_j y_i y_j \left(\mathbf{x}_i^T \mathbf{x}_j + \frac{\delta_{ij}}{C} \right)$$

$$= \sum_{ij} b_i b_j K_{ij} = \mathbf{b}^T \mathbf{K}\, \mathbf{b}$$

with

$$b_i = \alpha_i y_i; \quad K_{ij} = \mathbf{x}_i^T \mathbf{x}_j + \frac{\delta_{ij}}{C}$$

- this:
  - **increments** the eigenvalues by $1/C$, making the problem better conditioned
  - for larger C, the extra term is **smaller** and <u>**outliers**</u> have a **larger** influence (**less penalty** on them, **more reliance** on data term)

1−Norm SVM

$$\min_{\mathbf{w},\xi,b} \|\mathbf{w}\|^2 + C\sum_i \xi_i \quad \text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i$$
$$\xi_i \geq 0, \forall i$$

- and

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r}) =$$
$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i + \sum_i \alpha_i[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)] - \sum_i r_i\xi_i$$

- from which

$$\nabla_{\mathbf{w}} L = 0 \iff \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0 \iff \mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b L = 0 \iff \sum_i \alpha_i y_i = 0 \qquad \nabla_{\xi_i} L = 0 \iff C - \alpha_i - r_i = 0$$

1−Norm SVM

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i + \sum_i \alpha_i[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)] - \sum_i r_i \xi_i$$

- plugging back

$$\mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i \qquad \sum_i \alpha_i y_i = 0 \qquad r_i = C - \alpha_i$$

- we get the Lagrangian

$$L(\mathbf{w}^*, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r}) = \frac{1}{2}\sum_{ij} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \cancel{C\sum_i \xi_i} + \sum_i \alpha_i(1 - \cancel{\xi_i}) - \sum_{ij} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

$$-\underbrace{\sum_i \alpha_i y_i}_{0} - \sum_i (\cancel{C} - \cancel{\alpha_i})\xi_i$$

$$= -\frac{1}{2}\sum_{ij} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum_i \alpha_i$$

- this is exactly **like** the <u>hard−margin</u> case with the <u>extra</u> constraint $\alpha_i = C - r_i, \forall i$

Soft Dual for 1−Norm

extra constraint relative to hard−margin case

- in **summary**:

$$\boxed{\mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i} \quad \boxed{\sum_i \alpha_i y_i = 0} \quad \boxed{r_i = C - \alpha_i}$$

$$\boxed{L(\mathbf{w}^*, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) = -\tfrac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_i \alpha_i}$$

- Recall

$$\min_{\mathbf{w},\boldsymbol{\xi},b} \|\mathbf{w}\|^2 + C\sum_i \xi_i \quad \text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i$$
$$\xi_i \geq 0, \forall i$$

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}) = \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i + \boxed{\sum_i \alpha_i[1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)]} - \boxed{\sum_i r_i \xi_i}$$

- from the constraints

$$\boxed{y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \Leftrightarrow 1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b) \leq 0} \quad \boxed{\xi_i \geq 0 \Leftrightarrow -\xi_i \leq 0}$$

the KKT conditions are

$$\alpha_i > 0 \Leftrightarrow 1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b) = 0$$
$$\alpha_i = 0 \Leftrightarrow 1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b) < 0$$

$$r_i > 0 \Leftrightarrow \xi_i = 0$$
$$r_i = 0 \Leftrightarrow \xi_i > 0$$

Soft Dual for 1–Norm

extra constraint relative to hard–margin case

$$L(\mathbf{w}^*, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r}) = -\frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_i \alpha_i$$

$$\mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\sum_i \alpha_i y_i = 0$$

$$r_i = C - \alpha_i \quad (*)$$

a) $\alpha_i > 0 \Leftrightarrow 1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b) = 0$
b) $\alpha_i = 0 \Leftrightarrow 1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b) < 0$

c) $r_i > 0 \Leftrightarrow \xi_i = 0$
d) $r_i = 0 \Leftrightarrow \xi_i > 0$

- if $\boxed{\alpha_i = 0}$
  - from $(*)$, $r_i = C$ and, from c), $\xi_i = 0$
  - from b), $1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b) < 0$ and, since $\xi_i = 0$, we have $y_i(\mathbf{w}^T\mathbf{x}_i + b) > 1$, i.e. $\boxed{\mathbf{x}_i \text{ is correctly classified}}$
- if $\boxed{\alpha_i > 0}$
  - since r_i are Lagrange multipliers, $r_i \geq 0$, $(*)$ means that $\alpha_i \leq C$
  - if $r_i > 0 \Rightarrow \boxed{\alpha_i < C}$, from c) $\xi_i = 0$ and from a) $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$, i.e. $\boxed{\mathbf{x}_i \text{ is on the margin}}$
  - if $r_i = 0 \Rightarrow \boxed{\alpha_i = C}$, from d) $\xi_i > 0$ and from a) $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1 - \xi_i < 1$, i.e. $\boxed{\mathbf{x}_i \text{ is an outlier}}$

Soft Dual for 1−Norm

$\alpha_i = 0$, $\mathbf{x}_i$ is correctly classified

$0 < \alpha_i < C$, $\mathbf{x}_i$ is on the margin

$\alpha_i = C$, $\mathbf{x}_i$ is an outlier

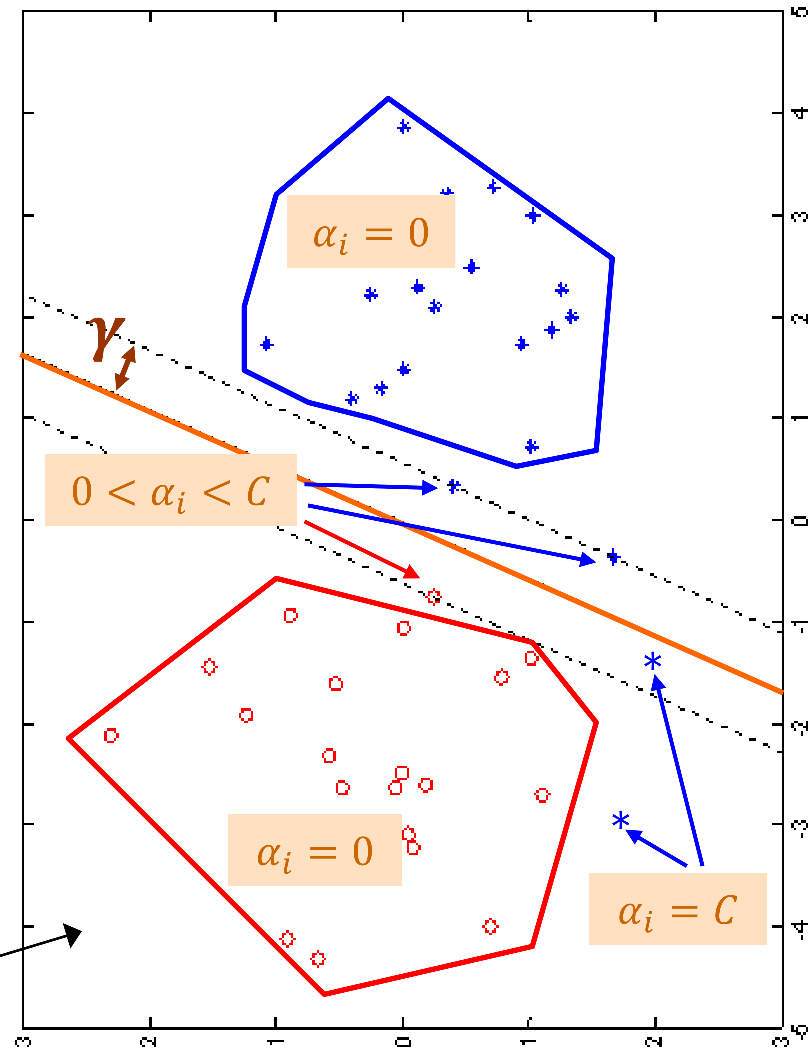
- overall, dual problem is

$$\max_{\boldsymbol{\alpha}\geq 0}\left\{-\frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum_i \alpha_i\right\}$$

$$\text{subject to } \sum_i \alpha_i y_i = 0,$$

$$0 \leq \alpha_i \leq C$$

- the only difference with respect to the hard−margin case is the "box constraint" on the α_i
- geometrically, we have this



Soft Dual for 1−Norm: Support Vectors

- support vectors are the points with $\alpha_i > 0$
- as before, the decision rule is

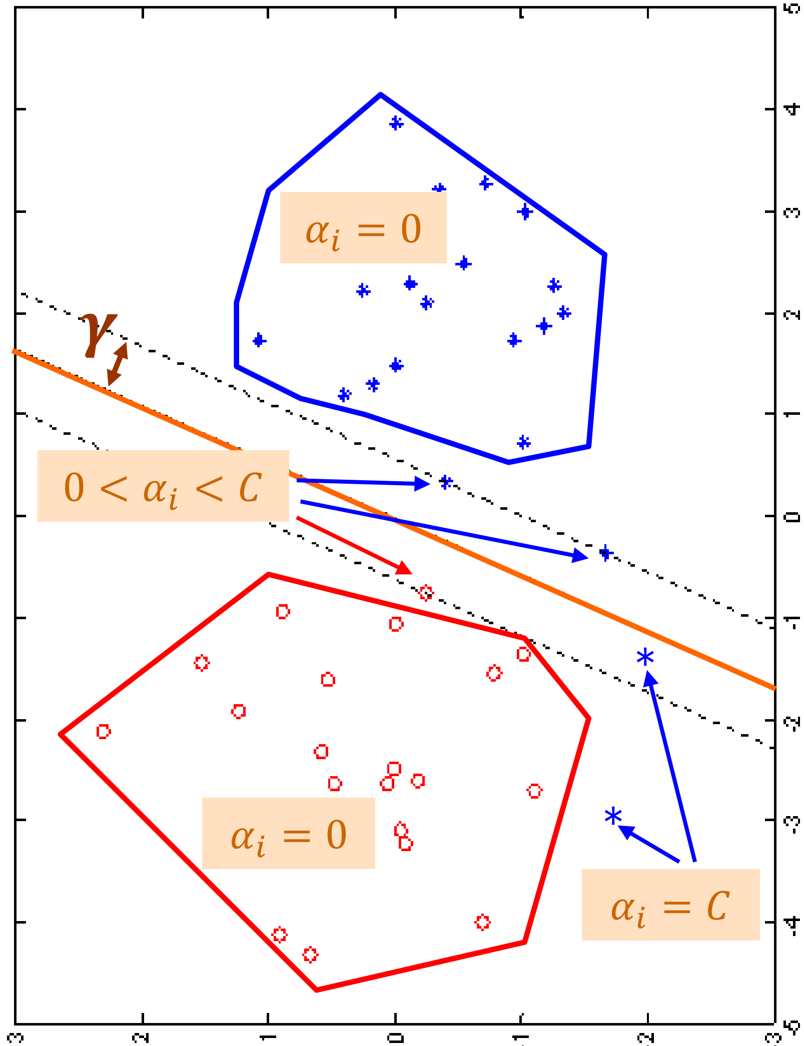
$$f(\mathbf{x}) = \operatorname{sgn}\left[\sum_{i \in SV} y_i \alpha_i^* \mathbf{x}_i^T \mathbf{x} + b^*\right]$$

  where $SV = \{i \mid \alpha_i^* > 0\}$ and b^* chosen such that

$$y_i g(\mathbf{x}_i) = 1, \forall \mathbf{x}_i \text{ s.t. } 0 < \alpha_i < C$$

- the box constraint on Lagrange multipliers makes **intuitive** sense:

  it prevents a single SV outlier from having large impact in the decision rule



Soft−Margin SVM

$$\min_{\mathbf{w},\xi,b} \ \|\mathbf{w}\|^2 + Cf(\xi)$$
$$\text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i$$
$$\xi_i \geq 0, \forall i$$

- note that C **controls** the importance of outliers
  - larger C implies that **more** emphasis is given to **minimizing the number of outliers**
- **1−norm** vs **2−norm**
  - as usual, the **1−norm** tends to **limit more** drastically the **outlier contributions**
  - this makes it a bit **more robust**, and it tends to be used more frequently in practice
- common problem:
  - **not** really intuitive **how to set up C**
  - usually **cross−validation**: there is a need to cross−validate with respect to both C and kernel parameters

υ – SVM

- an **alternative formulation** has been introduced to try to **overcome** this

$$\min_{\mathbf{w},\xi,\rho,b} \|\mathbf{w}\|^2 - \upsilon\rho + \frac{1}{n}\sum_i \xi_i$$

$$\text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq \rho - \xi_i, \forall i$$

$$\xi_i \geq 0, \forall i$$

$$\rho \geq 0, \forall i$$

- **advantages:**
  - υ has **intuitive** interpretation:

    1) υ is an upper bound on the proportion of training vectors that are margin errors, i.e. for which $y_i g(\mathbf{x}_i) \leq \rho$

    2) υ is a lower bound on total number of support vectors
  - more discussion on Quiz #4 (Prob. 3)

SVM: Connections to Regularization

- we talked about penalizing functions that are too complicated to improve **generalization**
- instead of the empirical risk, we should minimize the regularized risk

$$\boxed{R_{reg}[f] = R_{emp}[f] + \lambda\Omega[f]}$$

$$\min_{\mathbf{w},\xi,b} \;\; \|\mathbf{w}\|^2 + Cf(\xi)$$

$$\text{subject to} \;\; y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \forall i$$

$$\xi_i \geq 0, \forall i$$

- the SVM seems to be doing this in some sense:
  - it is designed to have as few errors as possible on **training set** (this is **controlled** by the **soft–margin weight** C)
  - we maximize **the margin** by **minimizing** $\|\mathbf{w}\|^2$ (which is a form of **complexity penalty**)
  - hence, **maximizing the margin** must be connected to enforcing some form of **regularization**

SVM: Connections to Regularization

- the connection can be made <u>explicit</u>
- consider the 1−norm SVM

$$\min_{\mathbf{w},\xi,b} \|\mathbf{w}\|^2 + C\sum_i \xi_i \quad \text{subject to} \quad \begin{array}{r} y_i g(\mathbf{x}_i) \geq 1 - \xi_i, \forall i \\ \xi_i \geq 0, \forall i \end{array}$$

- the constraints can be rewritten as

i) $\xi_i \geq 0$ and ii) $\xi_i \geq 1 - y_i g(\mathbf{x}_i)$

which is equivalent to

$$\boxed{\xi_i \geq \max[0, 1 - y_i g(\mathbf{x}_i)] = [1 - y_i g(\mathbf{x}_i)]_+}$$

- note that the cost $\|\mathbf{w}\|^2 + C\sum_i \xi_i$ can only **increase** with larger ξ_i
- hence, at the optimal solution, $\boxed{\xi_i^* = [1 - y_i g(\mathbf{x}_i)]_+}$

SVM: Connections to Regularization

- the problem

$\xi_i^* = \max[0, 1 - y_i g(\mathbf{x}_i)] = [1 - y_i g(\mathbf{x}_i)]_+$

$$\min_{\mathbf{w},b} \|\mathbf{w}\|^2 + C \sum_i [1 - y_i g(\mathbf{x}_i)]_+ \Leftrightarrow \boxed{\min_{\mathbf{w},b} \sum_i [1 - y_i g(\mathbf{x}_i)]_+ + \lambda \|\mathbf{w}\|^2}$$

(by making $\lambda = 1/C$)

can be seen as a

regularized risk

$$\boxed{R_{reg}[f] = \sum_i L[\mathbf{x}_i, y_i, f] + \lambda\Omega[f]}$$

with

- loss function $L[\mathbf{x}, y, g] = [1 - y g(\mathbf{x})]_+$ → hinge loss
- standard regularizer $\Omega[\mathbf{w}] = \|\mathbf{w}\|^2$

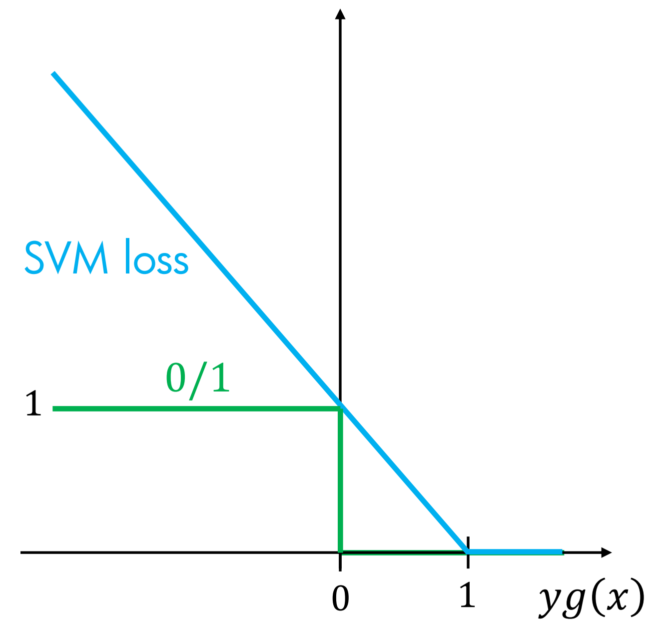
The SVM Loss

- it is interesting to **compare** the SVM loss

$$\boxed{L[\mathbf{x}, y, g] = [1 - yg(\mathbf{x})]_+}$$

with the $0/1$ loss:

- the SVM loss **penalizes large** negative margins
- assigns some penalty to anything with margin less than 1
- for the $0/1$ loss, the errors are **all** the same



- the **regularizer**

$$\boxed{\Omega[\mathbf{w}] = \|\mathbf{w}\|^2}$$

- **penalizes** planes of **large** $\|\mathbf{w}\|$
- standard **measure of complexity** in regularization theory

Recap: Risk Minimization

- note that **all** the methods we have studied minimize a similar risk

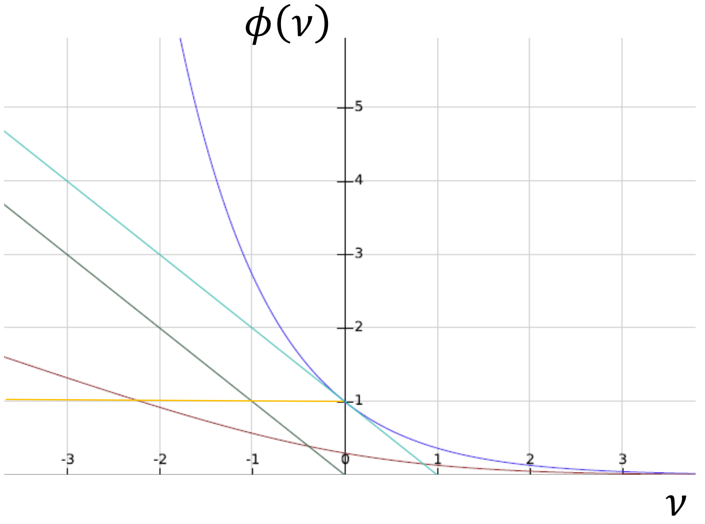
$$R_{reg}[f] = \sum_i L[y_i, f(\mathbf{x}_i)] + \lambda\Omega[f]$$

- in **all** cases, the loss function $L[y, g(\mathbf{x})]$ is a margin loss

$$L[y, g(\mathbf{x})] = \phi\big(yg(\mathbf{x})\big)$$

- only **difference** is the $\phi(\cdot)$ function

Method	$\phi(v)$
BDR (0/1 loss)	$\text{sign}(-v)$
Perceptron	$[-v]_+$
neural networks	$\log(1 + e^{-v})$
boosting	e^{-v}
SVM	$[1 - v]_+$



Recap: Risk Minimization

- note that **all** the methods we have studied minimize a **similar risk**

$$R_{reg}[f] = \sum_i L[y_i, f(\mathbf{x}_i)] + \lambda \Omega[f]$$

- the **regularizer** $\Omega[f]$ is implemented in **different** ways

Method	$\Omega[f]$
BDR (0/1 loss)	enforced in the estimation of the pdfs
Perceptron	none
neural networks	weight decays
boosting	regularization is implemented by limiting the number of iterations (weak learners)
SVM	$\|\|\mathbf{w}\|\|^2$

Recap: Risk Minimization

$$R_{reg}[f] = \sum_i L[y_i, f(\mathbf{x}_i)] + \lambda\Omega[f]$$

- note that minimizing R_{reg} is the same as maximizing

$$e^{-R_{reg}[f]} = e^{-\sum_i L[y_i, f(\mathbf{x}_i)]} \cdot e^{-\lambda\Omega[f]}$$

which is the **same** as

- finding the function f of **maximum** a posteriori probability
- under a probabilistic model with

likelihood function

$$e^{-\sum_i L[y_i, f(\mathbf{x}_i)]}$$

prior

$$e^{-\lambda\Omega[f]}$$

- hence, it has a **Bayesian interpretation**, where the **regularizer** defines the **prior**, which is used to constrain the values of the solution (e.g., if $\Omega[\mathbf{w}] = ||\mathbf{w}||^2$, the prior will be Gaussian with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$)

In Summary

- all **methods** are implementations of the same optimization framework
- loss functions can have significant difference (margin enforcing vs not)
- regularizers are more tied to the implementation