Maximal Margin

- The discussion motivates the notion of maximal margin
- The maximal margin of a data set S is defined as:

\[ \gamma(S) = \max_{||w||=1} \min_{(x,y) \in S} y w^T x \]

1. For a given w: Find the closest point.
2. Then, find the one that gives the maximal margin value across all w's (of size 1).

Note: the selection of the point in the min and therefore the max does not change if we scale w, so it's okay to only deal with normalized w's.
Hard SVM

• We want to choose the hyperplane that achieves the largest margin. That is, given a data set \( S \), find:

\[
\mathbf{w}^* = \text{argmax}_{||\mathbf{w}||=1} \min_{(x,y) \in S} y \mathbf{w}^T \mathbf{x}
\]

• **How to find this \( \mathbf{w}^* \)?**

• **Claim:** Define \( \mathbf{w}_0 \) to be the solution of the optimization problem:

\[
\mathbf{w}_0 = \text{argmin} \{ ||\mathbf{w}||^2 : \forall (x,y) \in S, y \mathbf{w}^T \mathbf{x} \geq 1 \}
\]

**Intuition:** “Freeze” the margin, minimize the weights

That is, the normalization of \( \mathbf{w}_0 \) corresponds to the largest margin separating hyperplane. Then:

\[
\frac{\mathbf{w}_0}{||\mathbf{w}_0||} = \text{argmax}_{||\mathbf{w}||=1} \min_{(x,y) \in S} y \mathbf{w}^T \mathbf{x}
\]

Hard SVM Optimization

• We have shown that the sought after weight vector \( \mathbf{w} \) is the solution of the following optimization problem:

**SVM Optimization:**

Minimize: \( \frac{1}{2} ||\mathbf{w}||^2 \)

Subject to: \( \forall (x,y) \forall S: y \mathbf{w}^T \mathbf{x} \geq 1 \)

• This is an optimization problem in \((n+1)\) variables, with \(|S|=m\) inequality constraints.
Support Vector Machines

- The name “Support Vector Machine” stems from the fact that $w^*$ is supported by (i.e., is the linear span of) the examples that are exactly at a distance $1 / ||w^*||$ from the separating hyperplane. These vectors are therefore called support vectors.

We just found the max margin classifier, by formulating the problem as a constrained minimization problem. **Are we done?**

### Non-Separable Case

Want to relax the constraints:

$$y_i(w \cdot x_i + b) \geq 1.$$

Introduce slack variables $\xi_i$:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i,$$

Where $\xi_i \geq 0$, an error occurs when $\xi_i > 0$

Thus we can assign an extra cost for errors, as follows:

Minimize

$$f(w, b, \xi) = \frac{1}{2}||w||^2 + C \sum_{i=1}^{m} \xi_i$$

subject to

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i; \quad \xi_i \geq 0, \quad i = 1, \ldots, m$$
Soft SVM

- Notice that the relaxation of the constraint: $y_i \mathbf{w}_i^\top \mathbf{x}_i \geq 1$ can be done by introducing a slack variable $\xi$ (per example) and requiring: $y_i \mathbf{w}_i^\top \mathbf{x}_i \geq 1 - \xi_i$; $\xi_i \geq 0$
- Now, we want to solve:
  \[
  \min \frac{1}{2} ||\mathbf{w}||^2 + c \sum \xi_i \quad \text{subject to } \xi_i \geq 0
  \]
- Which can be written as:
  \[
  \min \frac{1}{2} \mathbf{w}^\top \mathbf{w} + c \sum \max(0, 1 - y_i \mathbf{w}_i^\top \mathbf{x}_i).
  \]
Classification

Unfortunately, this is very difficult to minimize!
- Non convex and non differentiable

Instead, we will define a smooth approximation to the 0-1 loss, called a surrogate loss function.

Regression: Squared Loss

Is the square hinge loss a good candidate to be a surrogate loss function for 0-1 loss?
Surrogate Loss functions

- Surrogate loss function: smooth approximation to the 0-1 loss
  - Upper bound to 0-1 loss

![Graph of surrogate loss function]

Soft SVM (2)

- The hard SVM formulation assumes linearly separable data.
  - A natural relaxation: maximize the margin while minimizing the # of examples that violate the margin (separability) constraints.
  - This leads to non-convex problem that is hard to solve.
  - Instead, move to a surrogate loss function that is convex.

- SVM relies on the hinge loss function:
  \[
  \min_{w} \frac{1}{2} \|w\|^2 + c \max(0, 1 - y \cdot w^t x) \\
  \min_{w} \frac{1}{2} \|w\|^2 + c \sum_{(x,y) \in S} \max(0, 1 - y w^t x)
  \]
  - where the parameter $c$ controls the tradeoff between large margin (small $\|w\|$) and small hinge-loss.
SVM Objective Function

- The problem we solved is:  \(\text{Min } \frac{1}{2} ||w||^2 + c \sum \xi_i\)
- Where \(\xi_i > 0\) is a slack variable, defined as: \(\xi_i = \max(0, 1 - y_i w'x_i)\)
  - Equivalently, we can say that: \(y_i w'x_i \geq 1 - \xi_i\); \(\xi_i \geq 0\)
- And this can be written as:
  \[
  \text{Min } \frac{1}{2} ||w||^2 + c \sum \xi_i
  \]

- General Form of a learning algorithm:
  - Minimize empirical loss, and Regularize (to avoid over fitting)

What if the true model isn’t linearly separable?

- By transforming the feature space these functions can be made linearly separable
- Represent each point in 2D as \((x, x^2)\)
SVM: Kernel Function

- Don't need to transform dataset
  - Just need distance
- Kernel function: compute distance in new space without actually computing point positions!
  - Let $\phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$
  - $\phi(x)^T \phi(y) = x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2$
- Radial basis function: kernel representing infinite expansion
  - “Infinite dimension” feature space

Polynomial Regression

- GD: general optimization algorithm
  - Works for classification (different loss function)
  - Incorporate polynomial features to fit a complex model
  - Danger – overfitting!
Regularization

- Both for regression and classification, for a given error we prefer a simpler model
  - Keep $W$ small: $\epsilon$ changes in the input cause $\epsilon^*w$ in the output
- Sometimes we are even willing to trade a higher error rate for a simpler model (why?)
- Add a regularization term:
  - This is a form of inductive bias

$$\min_w = \sum_n \text{loss}(y_n, w_n) + \lambda R(w)$$

How different values impact learning?

Regularization

- A very popular choice of regularization term is to minimize the norm of the weight vector $||w|| = \sqrt{\sum_w w_i^2}$
  - For example $\min_w = \sum_n \text{loss}(y_n, w_n) + \frac{\lambda}{2} ||w||^2$
  - For convenience: $\frac{1}{2}$ squared norm
    - What is the update gradient of the loss function?
    - At each iteration we subtract the weights by $\lambda^*w$
- In general, we can pick other norms (p-norm)
  - Referenced as $L_p$ norm
  - Different $p$ will have a different effect!