Convex Optimization
SVM's and Kernel Machines

S.V.N. "Vishy" Vishwanathan

vishy@axiom.anu.edu.au

National ICT of Australia
and
Australian National University

Thanks to Alex Smola and Stéphane Canu

Overview

- Review of Convex Functions
- Convex Optimization
- Dual Problems
- Interior Point Methods
- Simple SVM
- Sequential Minimal Optimization (SMO)
- Miscellaneous Tricks of Trade

Convex Set and Function

Definition:

- A set $\mathfrak{X}$ (subset of a vector space) is convex iff

$$\lambda x + (1-\lambda)x' \in \mathfrak{X} \quad \forall x, x' \in \mathfrak{X} \text{ and } \lambda \in [0,1]$$

Convex Functions:

- A function $f : \mathfrak{X} \to \mathbb{R}$ is convex if for any $x, x' \in \mathfrak{X}$ and $\lambda \in [0,1]$ such that $\lambda x + (1-\lambda)x' \in \mathfrak{X}$

$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$$

- If strict inequality $\implies$ a strictly convex function

Below Sets:

- Let $f : \mathfrak{X} \to \mathbb{R}$ be convex. If $\mathfrak{X}$ is convex then the set $X := \{x \in \mathfrak{X} : f(x) \leq c\}$ is convex

Convex Minimization

Theorem:

- Function $f : \mathfrak{X} \to \mathbb{R}$ be convex
- The sets $\mathfrak{X}$ and $X \subseteq \mathfrak{X}$ be convex sets
- Let c be the minimum of f_X
- All $x \in X$ for which $f(x) = c$ form a convex set

Corollary:

- Let $f, c_1, c_2, \ldots c_n : \mathfrak{X} \to \mathbb{R}$ be convex
- The set $\mathfrak{X}$ be convex
- The optimization problem

$$\min_{x \in \mathfrak{X}} f(x)$$
$$\text{s.t } c_i(x) \leq 0$$

  has its solution (if it exists) on a convex set.
- If strictly convex functions solution is unique

Convex Maximization

Basic Idea:

- Convex maximization is generally hard
- Maximum attained on corner points or vertices

Maximization on an Interval:

- Let $f : [a, b] \longrightarrow \mathbb{R}$ be convex
- f attains its maximum at either a or b

Maxima of Convex Functions:

- Let X be a compact convex set in $\mathcal{X}$
- Denote by $|X$ the vertices of X
- Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be convex
- Then

$$\sup\{f(x)|x \in X\} = \sup\{f(x)|x \in |X\}$$

Newton Method

Basic Idea:

- We replace a function by its quadratic approximation
- If approximations are good $\implies$ fast convergence

Maximization on an Interval:

- Suppose $f : [a, b] \to \mathbb{R}$ is convex and *smooth*
- The following iterations converge to $\min f(x)$

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

Convergence:

- Let $g(x) := f'(x)$ be continuous and twice differentiable
- Let $x^* \in \mathbb{R}$ and $g(x^*) = 0$ and $g'(x^*) \neq 0$
- x_0 is sufficiently close to $x^* \implies$ quadratic convergence

Gradient Descent

Basic Idea:

- How do you climb up a hill?
- Take a step up and see how to go up again

Algorithm:

- Suppose $f : \mathcal{X} \to \mathbb{R}$ is convex and *smooth*
- The following iterations converge to $\min f(x)$

$$x_{n+1} = x_n - \gamma f'(x_n)$$

- Where $\gamma > 0$ minimizes f locally
- Such a γ must always exist

Convergence:

- Can show that converges in infinite steps
- We will not do the proof!

Predictor Corrector Methods

Basic Idea:

- Make a linear approximation to f
- Substitute your estimate into f and correct

Example:

- Suppose $f(x) = f_0 + ax + \frac{1}{2}bx^2$
- Linear approximation $f \approx f_0 + ax$
- Predictor solution $x_{pred} = -\frac{f_0}{a}$
- Substitute back $f_0 + ax_{corr} + \frac{1}{2}b(\frac{f_0}{a})^2$
- Solve $x_{corr} = -\frac{f_0}{a}\left(1 + \frac{1}{2}\frac{f_0 b}{a^2}\right)$
- Iterate until convergence
- Notice how we never compute $\sqrt{b}$!

KKT Conditions

Optimization:

- Optimization problem

$$\min f(x)$$
$$\text{s.t } c_i(x) \leq 0$$
$$e_j(x) = 0$$

Lagrange Function:

- Define

$$L(x, \alpha, \beta) := f(x) + \sum_{i=1}^{n} \alpha_i c_i(x) + \sum_{j=1}^{n'} \beta_j e_j(x)$$

$\alpha_i \geq 0$ and $\beta_j \in \mathbb{R}$

Theorem:

- If $L(\bar{\mathbf{x}}, \alpha, \beta) \leq L(\bar{\mathbf{x}}, \bar{\alpha}.\bar{\beta}) \leq L(x, \bar{\alpha}, \bar{\beta})$ then $\bar{\mathbf{x}}$ is a solution

Diff. Convex Problems

Optimization Problem:

- Optimization problem

$$\begin{aligned} \min f(x) \\ \text{s.t } c_i(x) \leq 0 \\ e_j(x) = 0 \end{aligned}$$

KKT Conditions:

- If f, c_i are convex and differentiable then $\bar{\mathbf{x}}$ is a solution if $\exists\, \bar{\alpha} \in \mathbb{R}^n$ s.t. $\alpha_i \geq 0$ and

$$\begin{aligned} \partial_x L(\bar{\mathbf{x}}, \bar{\alpha}) &= \partial_x f(\bar{\mathbf{x}}) + \sum_{i=1}^{n} \bar{\alpha}_i \partial_x c_i(\bar{\mathbf{x}}) = 0 \\ \partial_{\alpha_i} L(\bar{\mathbf{x}}, \bar{\alpha}) &= c_i(\bar{\mathbf{x}}) \leq 0 \\ \sum_i \bar{\alpha}_i c_i(x) &= 0 \end{aligned}$$

KKT Gap

Proximity to Solution:

- Let f and c_i be convex and differentiable
- For any (x, α) such that x feasible, $\alpha_i \geq 0$ and

$$\begin{aligned} \partial_x L(x, \alpha) &= 0 \\ \partial_{\alpha_i} L(x, \alpha) &\leq 0 \end{aligned}$$

- If $\bar{\mathrm{x}}$ is the optimal then the KKT gap is given by

$$f(x) \geq f(\bar{\mathbf{x}}) \geq f(x) + \sum_i \alpha_i c_i(x)$$

- Also called the duality gap

Duality:

- Instead of solving a primal problem solve a dual problem
- Find saddle point of $L(x, \alpha)$

Optimization Problem

Let $H_{ij} = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$, then

Primal Problem:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|\theta\|^2 \\ &\text{subject to} && y_i(\langle \theta, \mathbf{x}_i \rangle + b) - 1 \geq 0 \text{ for all } i \in \{1, 2, \ldots, m\} \end{aligned}$$

Dual Problem:

$$\begin{aligned} &\text{maximize} && -\frac{1}{2}\alpha^\top H \alpha + \sum_i \alpha_i \\ &\text{subject to} && \sum_i \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \text{ for all } i \in \{1, 2, \ldots, m\}. \end{aligned}$$

Generalized Dual Problem:

$$\begin{aligned} &\text{maximize} && -\frac{1}{2}\alpha^\top H \alpha + c^\top \alpha \\ &\text{subject to} && A\alpha = 0 \text{ and } 0 \leq \alpha_i \leq C \end{aligned}$$

SimpleSVM The Big Picture

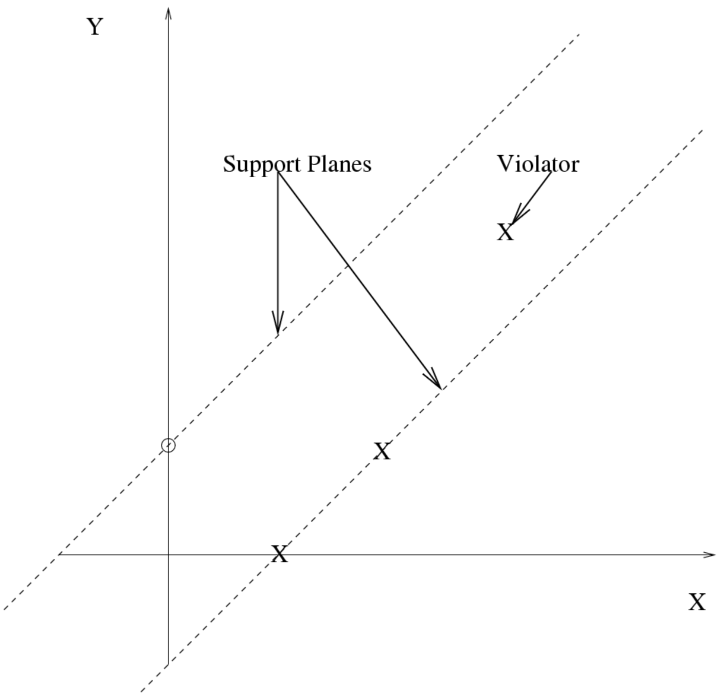
Chunking:

- Take chunks of data and solve the smaller problem.
- Retain the SV's, add the next chunk, and retrain.
- Repeat until convergence.

SimpleSVM:

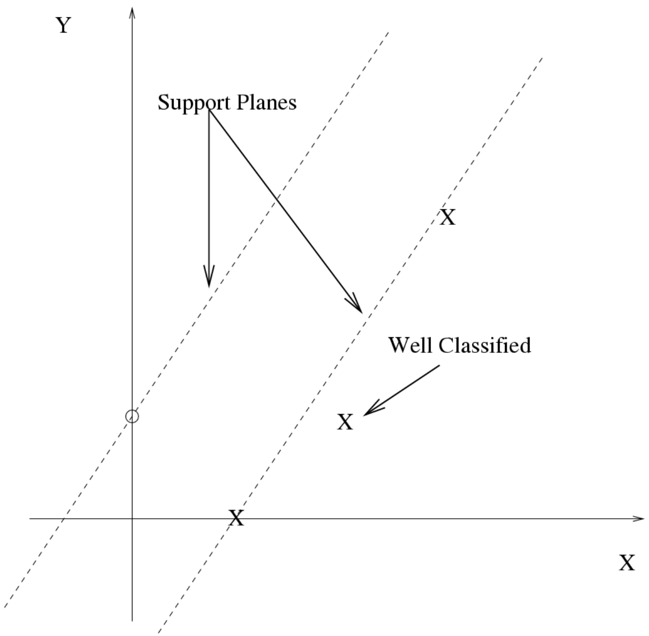
- Take an *active set* and optimize.
- Select a violating point greedily.
- Add violator to the active set.
- Add/delete only one SV at a time.
- Recompute the exact solution.
- Repeat until convergence.

A Picture Helps - I



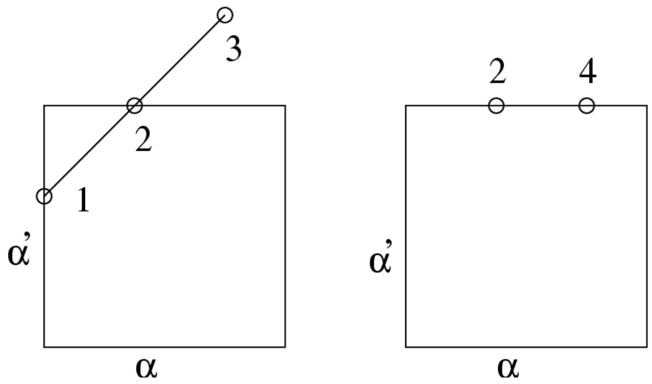
- Consider a hard margin linear SVM.
- x and o belong to different classes.
- Three SV's and one violator are shown.

A Picture Helps - II



- The support plane has been shifted.
- It passes thru the violator (rank-one update).
- A previous SV is now well classified (rank-one downdate).

What About Box Constraints?



- Point 1 is the current optimal solution.
- Add a new constraint α and optimize over (α, α').
- Point 3 is the unconstrained optimal.
- Move from point 3 to 2 where α' becomes bound.
- Now optimize over α to reach point 4.
- If 4 does not satisfy box constraints repeat.

Putting It Together

- Initialize with a *suitable* pair of points.
- Step 1:
  - Locate a violating point and add to the *active set*.
  - Ignore box constraints if any.
  - Solve the optimization problem for the active set.
  - Step 2:
    - If new solution satisfies box constraints we are done.
    - Else remove the first box constraint violator.
  - Goto Step 2.
  - Repeat until no violators (Goto Step 1).

Other Issues

Choosing Initial Points:

- Randomized strategies.
- Find a *good* pair with high probability.

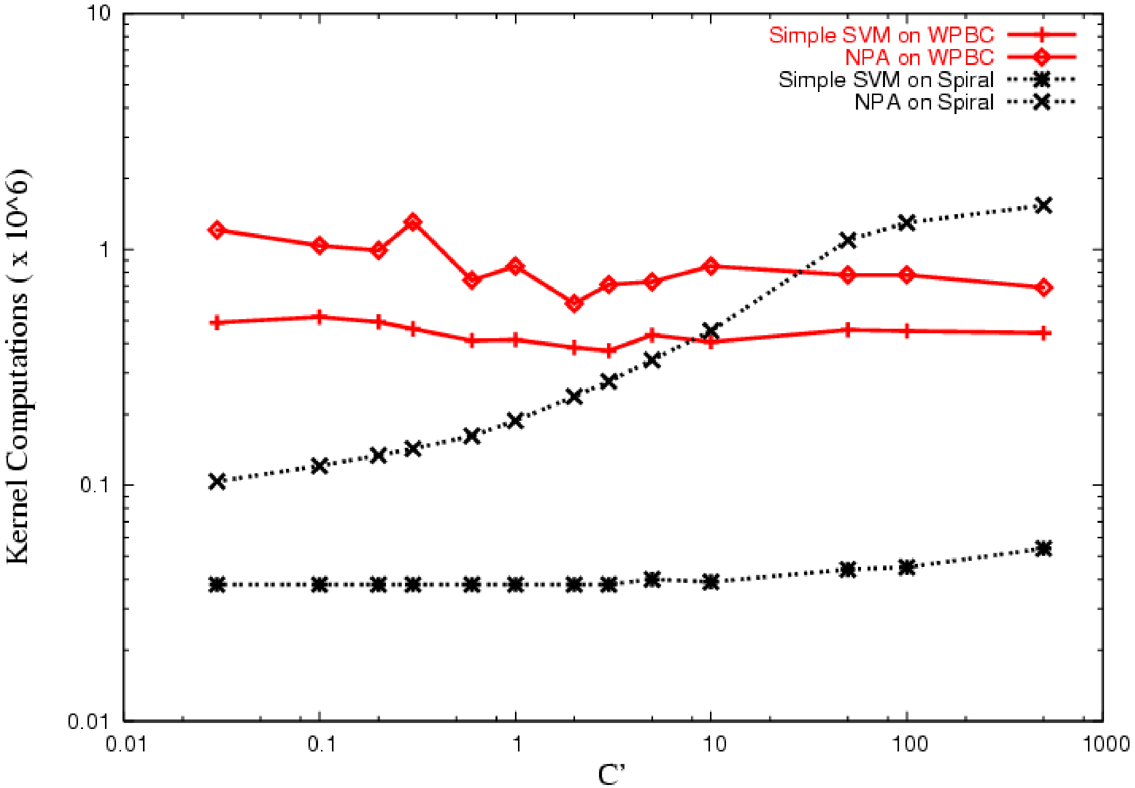
Rank One Updates:

- Kernel matrices are generally rank-degenerate.
- Cheap factorization algorithms.
- Cheap rank-one updates (Vishwanathan, 2002).

Convergence:

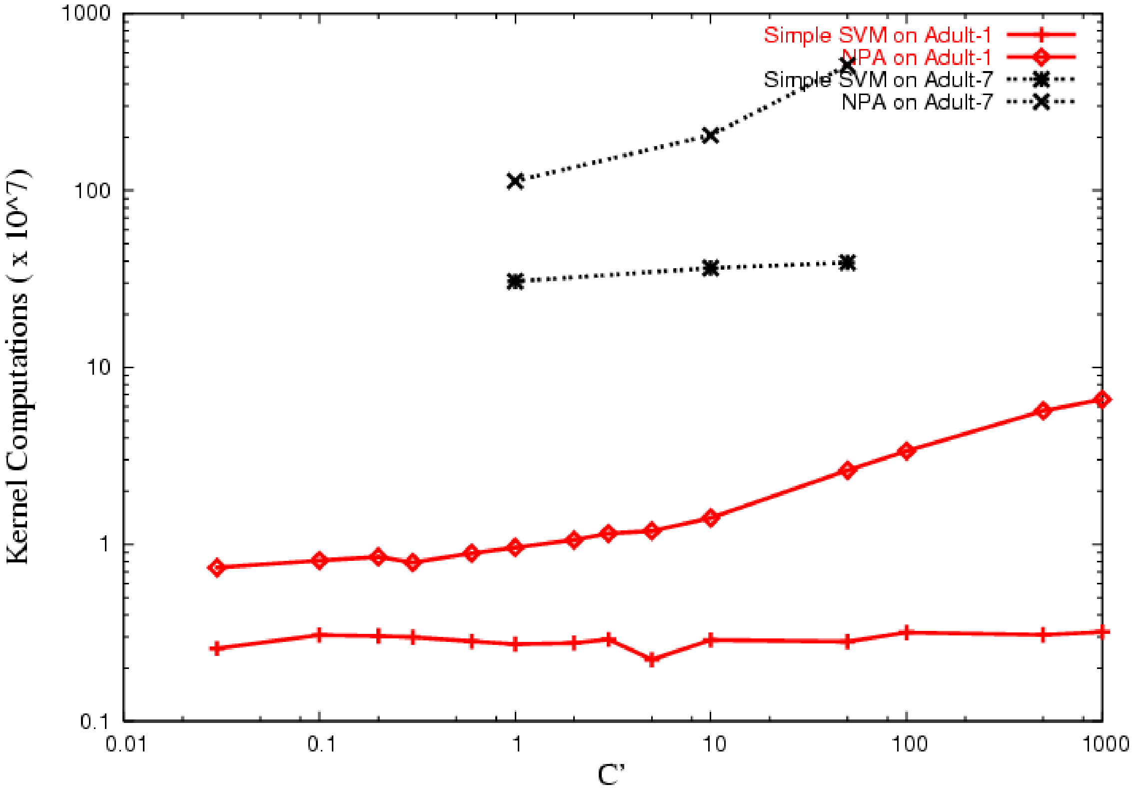
- Few sweeps thru the dataset suffice.
- Speed of convergence: Linear.

Performance Comparisons - I



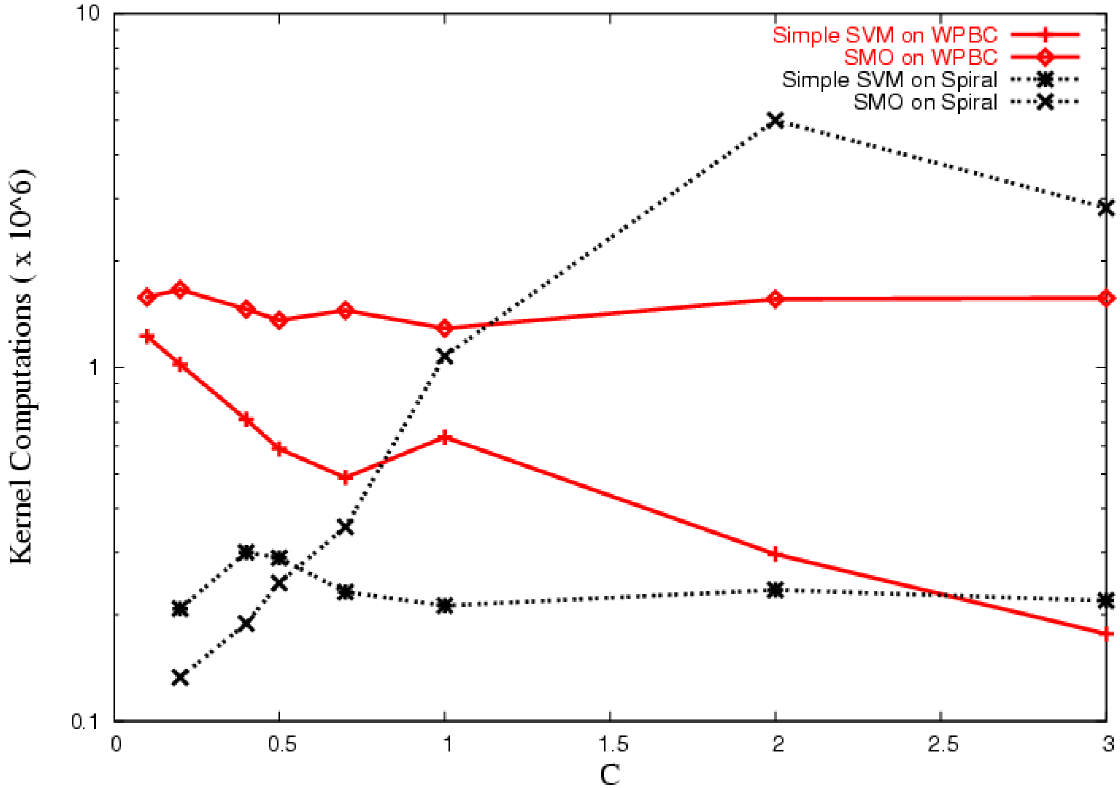
- Performance comparison between SimpleSVM and NPA on the *Spiral* dataset and the *WPBC* dataset.

Performance Comparisons - II



- Performance comparison between SimpleSVM and NPA on the *Adult* dataset.

Performance Comparisons - III



- Performance comparison between SimpleSVM and SMO on the *Spiral* and *WPBC* datasets.

Questions?