Approximating Concavely Parameterized Optimization Problems

Joachim Giesen  Sören Laue  Jens Mueller  Sascha Swiercy

Friedrich-Schiller-Universität Jena, Germany

NIPS 2012
Where is Jena?
Introduction

This talk is about general regularization paths.

- combinatorial complexity
- algorithms

Generalize and Simplify
General Regularization Paths

two forms

\[
\begin{align*}
\min_x & \quad \text{loss}(x) \\
\text{s. t.} & \quad \text{reg}(x) \leq t
\end{align*}
\]

\[
\begin{align*}
\min_x & \quad \text{reg}(x) + t \cdot \text{loss}(x)
\end{align*}
\]

examples

dual SVM

\[
\begin{align*}
\max_\alpha & \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s. t.} & \quad y^T \alpha = 0 \\
& \quad 0 \leq \alpha \leq t
\end{align*}
\]

primal SVM

\[
\begin{align*}
\min_w & \quad \|w\|_2^2 + t \cdot \text{loss}(w, X)
\end{align*}
\]
Problem: How to find the best regularization parameter $t$?
Regularization Paths

compute exact regularization path

▶ can be piecewise linear
▶ can have exponential combinatorial complexity
▶ can be numerically instable

in practice

▶ grid search
▶ resort to approximate regularization path
Abstract Problem

approximate regularization paths

Problem:

\[ f(t) := \min_x g(x; t) \]
Grid Search / General Idea
Grid Search / General Idea

grid search
Grid Search / General Idea

grid search
Grid Search / General Idea

grid search
Grid Search / General Idea

grid search
Grid Search / General Idea

grid search
Grid Search / General Idea

- not adaptive
- interesting regions and uninteresting regions treated the same

grid search
Grid Search / General Idea

grid search

adaptive step size
Grid Search / General Idea

grid search

adaptive step size
Grid Search / General Idea

grid search

adaptive step size
Grid Search / General Idea

grid search

adaptive step size
Grid Search / General Idea

- Grid search
- Adaptive step size
Grid Search / General Idea

grid search

adaptive step size
Grid Search / General Idea

grid search

adaptive step size
Grid Search / General Idea

grid search

adaptive step size
Grid Search / General Idea

grid search

adaptive step size
Grid Search / General Idea

grid search

adaptiver step size
Grid Search / General Idea

- Grid search
- Adaptive step size

The diagrams illustrate the concept of grid search and adaptive step size in the context of optimization or numerical methods.
Grid Search / General Idea

grid search

adaptive step size
Grid Search / General Idea

grid search

adaptive step size
Grid Search / General Idea

grid search

\sim Riemann integral

adaptive step size

\sim Lesbegue integral
Adaptive Step Size

\[ t_{\min} \leq t \leq t_{\max} \]

- cover whole path with \( \varepsilon \)-guarantee
- need \( \Theta(\frac{1}{\varepsilon}) \) many constant solutions
Adaptive Step Size

- cover whole path with $\varepsilon$-guarantee
Adaptive Step Size

- cover whole path with $\varepsilon$-guarantee
- need $\Theta(1/\varepsilon)$ many constant solutions
Adaptive Step Size – Constrained Form

apply this idea to

$$\min_x \text{ loss}(x)$$
$$\text{s. t. } \text{reg}(x) \leq t$$
Path Following Algorithm

dual SVM

\[
\begin{align*}
\max & \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s. t.} & \quad y^T \alpha = 0 \\
& \quad 0 \leq \alpha \leq t \\
\end{align*}
\]

\[=: f(t)\]
Path Following Algorithm

dual SVM

\[
\begin{align*}
\max \quad & -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} \quad & y^T \alpha = 0 \\
& 0 \leq \alpha \leq t \\
\end{align*}
\]

\[=: f(t)\]
Path Following Algorithm

dual SVM

\[
\begin{align*}
\max \quad & -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} \quad & y^T \alpha = 0 \\
& 0 \leq \alpha \leq t \\
\end{align*}
\]

\[= f(t)\]
Path Following Algorithm

\begin{align*}
\text{dual SVM} \\
& \quad \max -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
& \quad \text{s.t.} \quad y^T \alpha = 0 \\
& \quad \quad 0 \leq \alpha \leq t \\
& =: f(t)
\end{align*}
Path Following Algorithm

\[ t = t_1 \]

**dual SVM**

\[
\text{max } -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t. } y^T \alpha = 0 \\
0 \leq \alpha \leq t \\
\]

\[ =: f(t) \]
Path Following Algorithm

$$\begin{align*}
\max & \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} & \quad y^T \alpha = 0 \\
& \quad 0 \leq \alpha \leq t \\
& \quad =: f(t)
\end{align*}$$
Path Following Algorithm

dual SVM

\[\begin{align*}
\max & \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} & \quad \gamma^T \alpha = 0 \\
& \quad 0 \leq \alpha \leq t
\end{align*}\]

\[= f(t)\]
Path Following Algorithm

\[ \epsilon \]

\[ t_1 \quad t_2 \]

\[ t \]

dual SVM

\[
\begin{align*}
\max \quad & -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} \quad & y^T \alpha = 0 \\
& 0 \leq \alpha \leq t
\end{align*}
\]

\[ =: f(t) \]
Path Following Algorithm

dual SVM

\[
\begin{align*}
\max & \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} & \quad y^T \alpha = 0 \\
& \quad 0 \leq \alpha \leq t
\end{align*}
\]

\[=: f(t)\]
Path Following Algorithm

dual SVM

$$\max \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha$$

s. t.  
$$y^T \alpha = 0$$
$$0 \leq \alpha \leq t$$

$$=: f(t)$$
Path Following Algorithm

\[ \begin{align*}
\text{dual SVM} \\
\max & \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} & \quad y^T \alpha = 0 \\
& \quad 0 \leq \alpha \leq t \\
\end{align*} \]

\[ =: f(t) \]
Path Following Algorithm

\[
\begin{align*}
\text{dual SVM} \\
\max & \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} & \quad y^T \alpha = 0 \\
& \quad 0 \leq \alpha \leq t \\
\therefore & \quad f(t)
\end{align*}
\]
Path Following Algorithm

dual SVM

\[
\max \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} \quad y^T \alpha = 0 \\
0 \leq \alpha \leq t
\]

\[=: f(t)\]
Path Following Algorithm

dual SVM

\[
\max \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} \quad y^T \alpha = 0 \\
\quad 0 \leq \alpha \leq t
\]

\[=: f(t)\]
Path Following Algorithm

dual SVM

\[
\max \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} \quad y^T \alpha = 0 \\
0 \leq \alpha \leq t
\]

\[=: f(t)\]
Path Following Algorithm

dual SVM

\[
\max \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s. t.} \quad y^T \alpha = 0 \\
\quad 0 \leq \alpha \leq t
\]

\[=: f(t)\]
Path Following Algorithm

dual SVM

\[
\begin{align*}
\text{max} & \quad -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} & \quad y^T \alpha = 0 \\
& \quad 0 \leq \alpha \leq t \\
\end{align*}
\]

\[=: f(t)\]
Path Following Algorithm

dual SVM

\[
\begin{align*}
\max & \quad -\frac{1}{2}\alpha^T K\alpha + 1^T \alpha \\
\text{s.t.} & \quad y^T \alpha = 0 \\
& \quad 0 \leq \alpha \leq t \\
\end{align*}
\]

\[=: f(t)\]
Path Following Algorithm

dual SVM

\[
\max -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} \quad y^T \alpha = 0 \\
\quad 0 \leq \alpha \leq t \\
\]

\[=: f(t)\]
Path Following Algorithm

dual SVM

\[
\max -\frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s. t.} \quad y^T \alpha = 0 \\
0 \leq \alpha \leq t \\
=: f(t)
\]
Path Following Algorithm

dual SVM

\[
\begin{align*}
\max & \quad \frac{1}{2} \alpha^T K \alpha + 1^T \alpha \\
\text{s.t.} & \quad y^T \alpha = 0 \\
& \quad 0 \leq \alpha \leq t \\
\end{align*}
\]

\[=: f(t)\]
General Problem

\[ \min_x \text{ loss}(x) \]
\[ \text{s.t.} \quad \text{reg}(x) \leq t \]

- cover path with \( \varepsilon \)-guarantee
- need \( \Theta(1/\varepsilon) \) many constant solutions

works for any convex machine learning problem (e.g. kernelized SVMs, SDPs for matrix completion)
Adaptive Step Size – Additive Form

apply same idea to

$$\min_x \text{reg}(x) + t \cdot \text{loss}(x)$$
Path Following Algorithm

primal SVM

$$\min_w \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t)$$
Path Following Algorithm

\[ \min_w \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t) \]
Path Following Algorithm

primal SVM

$$\min_w \|w\|_2^2 + t \cdot \text{loss}(w, X)$$

$$= f(t)$$
Path Following Algorithm

primal SVM

$$\min_w \|w\|^2_2 + t \cdot \text{loss}(w, X)$$

$$= f(t)$$
Path Following Algorithm

primal SVM

$$\min_w \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t)$$
Path Following Algorithm

\[ \min_{w} \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t) \]
Path Following Algorithm

primal SVM

$$\min_w \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t)$$
Path Following Algorithm

primal SVM

\[
\min_w \|w\|^2_2 + t \cdot \text{loss}(w, X) =: f(t)
\]
Path Following Algorithm

\[
\text{primal SVM} \quad \min_w \quad \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t)
\]
Path Following Algorithm

primal SVM

\[ \min_w \|w\|^2_2 + t \cdot \text{loss}(w, X) =: f(t) \]

\[ t_1 \quad \bullet \quad t_2 \quad \rightarrow \quad t \]
Path Following Algorithm

primal SVM

\[
\min_w \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t)
\]
Path Following Algorithm

primal SVM

$$\min_w \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t)$$
Path Following Algorithm

Primal SVM

\[
\min_w \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t)
\]
Path Following Algorithm

primal SVM

$$\min_w \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t)$$
Path Following Algorithm

\[
\begin{align*}
\min_w &\quad \|w\|_2^2 + t \cdot \text{loss}(w, X) \\
= &\quad f(t)
\end{align*}
\]
Path Following Algorithm

primal SVM

$$\min_w \|w\|^2_2 + t \cdot \text{loss}(w, X)$$

$$= f(t)$$
Path Following Algorithm

primal SVM

\[
\min_w \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t)
\]
Path Following Algorithm

primal SVM

\[
\min_w \|w\|_2^2 + t \cdot \text{loss}(w, X) =: f(t)
\]
General Problem

\[
\min_x \ reg(x) + t \cdot \text{loss}(x)
\]

- cover path with \( \varepsilon \)-guarantee
- need \( \Theta(1/\sqrt{\varepsilon}) \) many constant solutions

works for any convex machine learning problem (e.g. kernelized SVMs, LASSO, SDPs for matrix completion)
Path Algorithm

Algorithm:

- $t = t_{\text{min}}$
- compute $\frac{\varepsilon}{2}$-approximate solution $x^*$ for parameter $t$
- compute next $t$ such that $x^*$ stays $\varepsilon$-approximate solution
- repeat until $t = t_{\text{max}}$
Regularization Paths / Cross-validation

Support Vector Machine Classification
Regularization Paths / Cross-validation

Nuclear Norm Regularized Matrix Completion

A Hybrid Algorithm for Convex Semidefinite Optimization (ICML2012)
(very fast SDP solver)
Path Complexity

path complexity vs $\varepsilon$
Conclusion — Generalize and Simplify

- general regularization path method
- works for SVMs, LASSO, matrix completion, SDPs, ...
- piecewise constant path, with guarantees
- $\Theta(1/\varepsilon)$ complexity for
  \[
  \min_x \text{ loss}(x) \\
  \text{s. t.} \quad \text{reg}(x) \leq t
  \]

- $\Theta(1/\sqrt{\varepsilon})$ complexity for
  \[
  \min_x \text{ reg}(x) + t \cdot \text{loss}(x)
  \]