Bundle Methods for Machine Learning

Joint work with Quoc Le, Choon-Hui Teo, Vishy Vishwanathan and Markus Weimer

Alexander J. Smola

sml.nicta.com.au
Statistical Machine Learning Program
Canberra, ACT 0200 Australia
Alex.Smola@nicta.com.au

Tokyo, October 12, 2007

Outline

Data

Observations

- Images
- Strings
- Movie rentals logs and scores
- Webpages
- Microarray measurements

Labels

- Identity of users, objects, biometric features
- Named entities, tags, paragraph segmentation
- Lists of preferred movies, related entities
- Ranking
- Health status, relevance of genes

Loss

Sophisticated discrepancy score for estimated label.

Loss Functions

Example: Density estimation in exponential families

- Find maximizer of log-likelihood

$$-\log p(y|x) = \log \sum_{y'} e^{f(x,y')} - f(x,y)$$

Example: Winner takes all estimation

- Estimate label $y^*(x)$ for observation x via

$$y^*(x) = \operatorname*{argmax}_y f(x,y) \text{ and incur loss } \Delta(y, y^*(x)).$$

- This problem is nonconvex in f. Convex bound via

$$\Delta(y, y^*(x)) \leq \max_{y'} f(x,y') - f(x,y) + \Delta(y,y')$$

Example: Least Mean Squares Regression

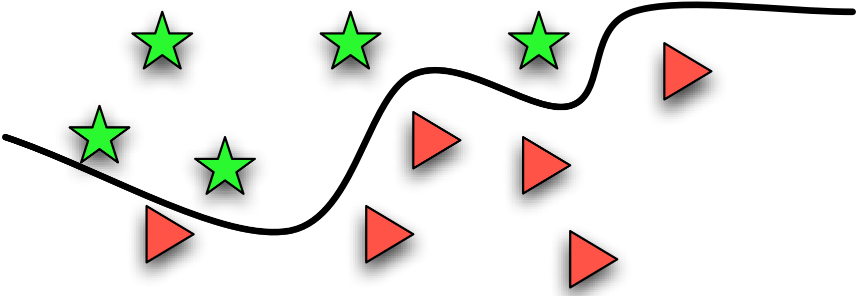
Binary Classification

Decision Function

$$f(x, y) = yf(x) \text{ where } y \in \{\pm 1\}$$

Estimate

$$y^*(x) = \operatorname*{argmax}_{y \in \{\pm 1\}} yf(x) = \operatorname{sgn} f(x)$$



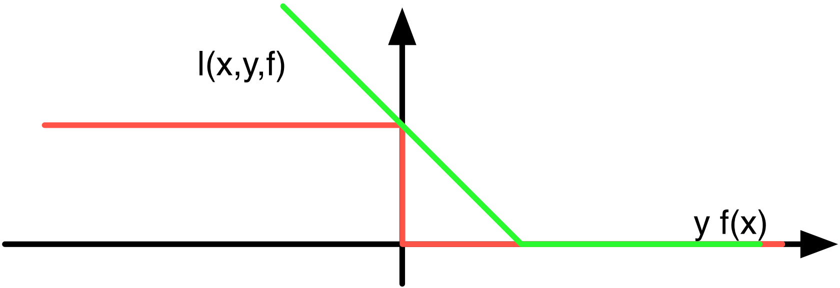
Binary Classification

Loss Function

$$\Delta(y, y') = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{otherwise} \end{cases}$$

Convex Upper Bound (soft margin loss)

$$l(x, y, f) = \max(0, 1 - yf(x))$$



Segmentation

Paragraph Segmentation

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, ‘and what is the use of a book,’ thought Alice ‘without pictures or conversation?’

<break>

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

<break>

There was nothing so very remarkable in that; nor did Alice think it so very much out of the way to hear the Rabbit say to itself, ‘Oh dear! Oh dear! I shall be late!’ (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually took a watch out of its waistcoat-pocket, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.

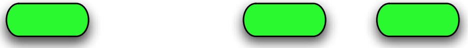
<break>

In another moment down went Alice after it, never once considering how in the world she was to get out again.

<break>

The rabbit-hole went straight on like a tunnel for some way, and then dipped suddenly down, so suddenly that Alice had not a moment to think about stopping herself before she found herself falling down a very deep well.

Protein Positioning



GATTACATTACTCAGTACTCAGGTCTCTATCTGATTACATTACTCAGTACTCAGGTCTCTATCT

Segmentation

Labels

$y = \{1, 5, 23, 49, 99, \ldots\}$ is a list of positions, i.e. $y \subset \{1, \ldots, n\}$.

Loss

1. Unit loss for each missed and each wrongly placed segment boundary.
2. Increasing loss for wrongly placed boundaries.

The Argmax

The function $f(x, y)$ has the semi Markov property.

$$f(x, y) = \sum_i \bar{f}(x, y_i, y_{i+1}, y_{i+2})$$

Maximize it by dynamic programming. Note that the number of segments need not be fixed.

Web Page Ranking

Top ranking Google scores for "euro 2007"

1. 22nd European Conference on Operational Research
2. Live Score service (powered by LiveScore.com)
3. CAP Euro 2007 - October 4 - 7th Barcelona, Spain
4. Under-21 squad readies their Euro 2007 finals campaign
5. Euro-Par 2007 Conference in Rennes

Discounted Cumulative Gains Score

Find a permutation π such that for ratings y_i we maximize

$$\mathrm{DCG}(y, \pi) = \sum_i \frac{2^{y_{\pi(i)}}}{\log(i+1)}$$

The Argmax function

$$f(x, \pi) = \sum_i c_{\pi(i)} \bar{f}(x_i) \text{ is maximized by sorting.}$$

Linear Function Classes

Key Observation

Many loss functions can be made convex in f.

Consequences

- Only useful if f is chosen from a vector space.
- Use Banach spaces
- Reproducing Kernel Hilbert Spaces are powerful since

$$\langle f, k(x, \cdot)\rangle = f(x)$$

Representer theorems and parametric problems.

Simplified Representation

$f(x, y) = \langle \phi(x, y), w\rangle$ for some feature map $\phi(x, y)$.

Regularized Risk Functional

Empirical Risk

$$R_{\mathrm{emp}}[w] = \frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, w) \text{ where } l \text{ is a convex loss.}$$

Applications include classification, regression, quantile regression, ranking, segmentation, sequence annotation, named entity tagging, Poisson, ...

Overfitting

Add regularizer to $R_{\mathrm{emp}}[w]$ and minimize $R_{\mathrm{emp}}[w] + \lambda\Omega[w]$.

Regularizers

- Quadratic regularization $\Omega[w] = \frac{1}{2}\|w\|_2^2$.
- LP regularization $\Omega[w] = \frac{1}{2}\|w\|_1^2$.
- Entropy regularization $\Omega[w] = \sum_i w_i \log w_i$.

The Chinese Restaurant guide to writing machine learning papers

Step 1: pick a loss function $l(x, y, w)$
Bonus points if you find with a new one.

Step 2: pick a regularizer $\Omega[w]$
Bonus points if you find with a new one (happens rarely).

Step 3: pick a new feature map
Bonus points if you can compute $\langle \phi(x, y), w \rangle$ cheaply.

Step 4: build a fancy implementation
Must run faster on at least one problem.

Publication
Happens if at least one of the four features is new.

A better idea

One Algorithm to rule them all, One Algorithm to find them, One Algorithm to bring them all and in the darkness bind them . . .

Outline

Key Idea

Empirical Risk

- Convex
- Expensive to compute
- Line search just as expensive as new computation
- Gradient comes almost for free with function value
- Parallel computation simple

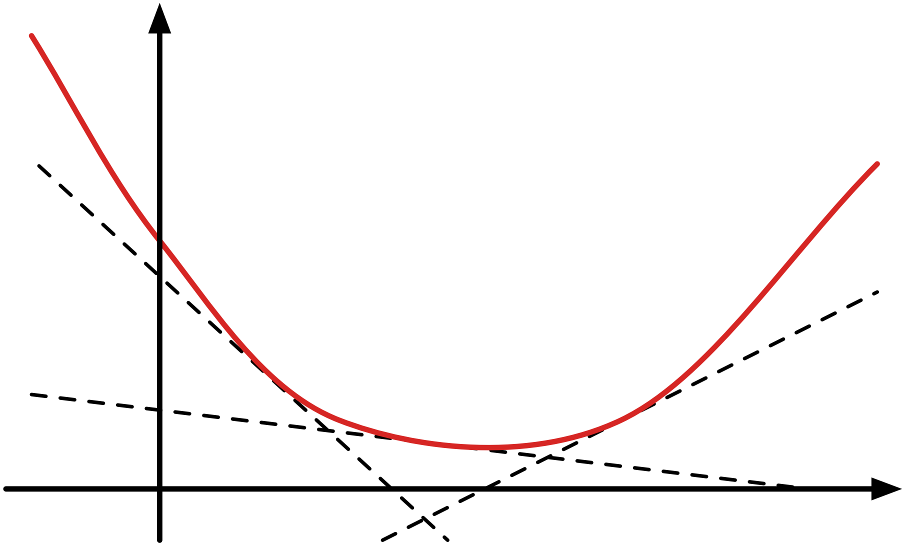
Regularizer

- Convex
- Cheap to compute
- Cheap to optimize

Strategy

- Compute only tangents on emprirical risk
- Perform optimization in the dual
- Modularity

Bundle Approximation



Lower Bound

Regularized Risk Minimization

$$\underset{w}{\text{minimize}}\, R_{\text{emp}}[w] + \lambda\Omega[w]$$

Taylor Approximation for $R_{\text{emp}}[w]$

$$R_{\text{emp}}[w] \geq R_{\text{emp}}[w_t] + \langle w - w_t, \partial_w R_{\text{emp}}[w_t]\rangle = \langle a_t, w\rangle + b_t$$

where $a_t = \partial_w R_{\text{emp}}[w_{t-1}]$ and $b_t = R_{\text{emp}}[w_{t-1}] - \langle a_t, w_{t-1}\rangle$.

Bundle Bound

$$R_{\text{emp}}[w] \geq R_t[w] := \max_{i \leq t} \langle a_i, w\rangle + b_i$$

Regularizer $\Omega[w]$ solves stability problems.

Algorithm

Pseudocode

Initialize $t = 0, w_0 = 0, a_0 = 0, b_0 = 0$
repeat
  Find minimizer

$$w_t := \operatorname*{argmin}_w R_t(w) + \lambda\Omega[w]$$

  Compute gradient a_{t+1} and offset b_{t+1}.
  Increment $t \leftarrow t + 1$.
until $\epsilon_t \leq \epsilon$

Convergence Monitor $R_{t+1}[w_t] - R_t[w_t]$

Since $R_{t+1}[w_t] = R_{\mathrm{emp}}[w_t]$ (Taylor approximation) we have

$$R_{t+1}[w_t] + \lambda\Omega[w_t] \geq \min_w R_{\mathrm{emp}}[w] + \lambda\Omega[w] \geq R_t[w_t] + \lambda\Omega[w_t]$$

Dual Problem

Good News

Dual optimization for $\Omega[w] = \frac{1}{2}\|w\|_2^2$ is Quadratic Program regardless of the choice of the empirical risk $R_{\mathrm{emp}}[w]$.

Details

$$\underset{\beta}{\text{minimize}}\ \frac{1}{2\lambda}\beta^\top AA^\top\beta - \beta^\top b$$

$$\text{subject to } \beta_i \geq 0 \text{ and } \|\beta\|_1 = 1$$

The primal coefficient w is given by $w = -\lambda^{-1}A^\top\beta$.

General Result

Use Fenchel-Legendre dual of $\Omega[w]$, e.g. $\|\cdot\|_1 \to \|\cdot\|_\infty$.

Very Cheap Variant

Can even use simple line search for update (almost as good).

Features

Parallelization

- Empirical risk sum of many terms: MapReduce
- Gradient sum of many terms, gather from cluster.
- Possible even for multivariate performance scores.

Solver independent of loss

No need to change solver for new loss.

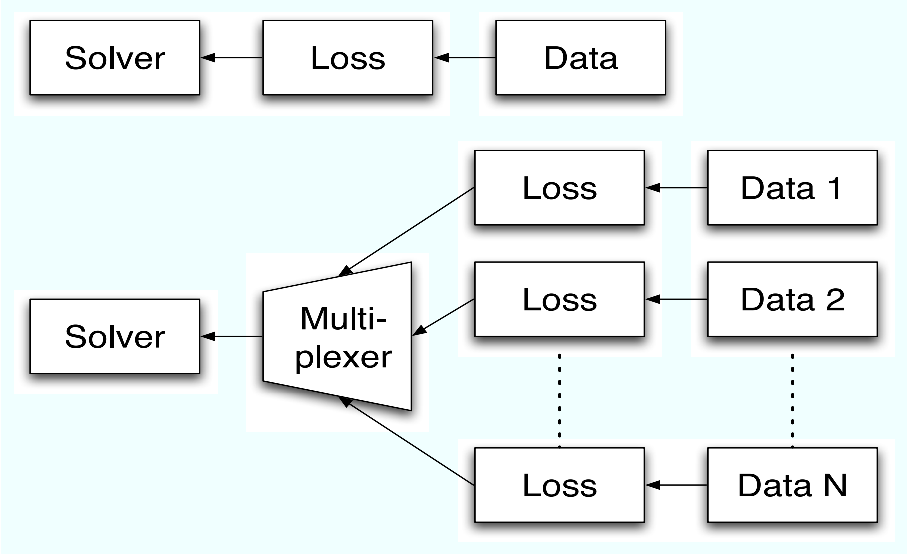
Loss independent of solver/regularizer

Add new regularizer without need to re-implement loss.

Line search variant

- Optimization does not require QP solver at all!
- Update along gradient direction in the dual.
- We only need inner product on gradients!

Architecture



Outline

Convergence

Theorem

The number of iterations to reach ϵ precision is bounded by

$$n \leq \log_2 \frac{\lambda R_{\mathrm{emp}}[0]}{G^2} + \frac{8G^2}{\lambda\epsilon} - 4$$

steps. If the Hessian of $R_{\mathrm{emp}}[w]$ is bounded, convergence to any $\epsilon \leq \lambda/2$ takes at most the following number of steps:

$$n \leq \log_2 \frac{\lambda R_{\mathrm{emp}}[0]}{4G^2} + \frac{4}{\lambda} \max\left[0, 1 - 8G^2 H^*/\lambda\right] - \frac{4H^*}{\lambda} \log 2\epsilon$$

Advantages

- Linear convergence for smooth loss
- For non-smooth loss almost as good in practice (as long as smooth on a course scale).
- Does not require primal line search.

Proof Idea

Duality Argument

- Dual of $R_i[w] + \lambda\Omega[w]$ lower bounds minimum of regularized risk $R_{\mathrm{emp}}[w] + \lambda\Omega[w]$.
- $R_{i+1}[w_i] + \lambda\Omega[w_i]$ is upper bound.
- **Show that the gap $\gamma_i := R_{i+1}[w_i] - R_i[w_i]$ vanishes.**

Dual Improvement

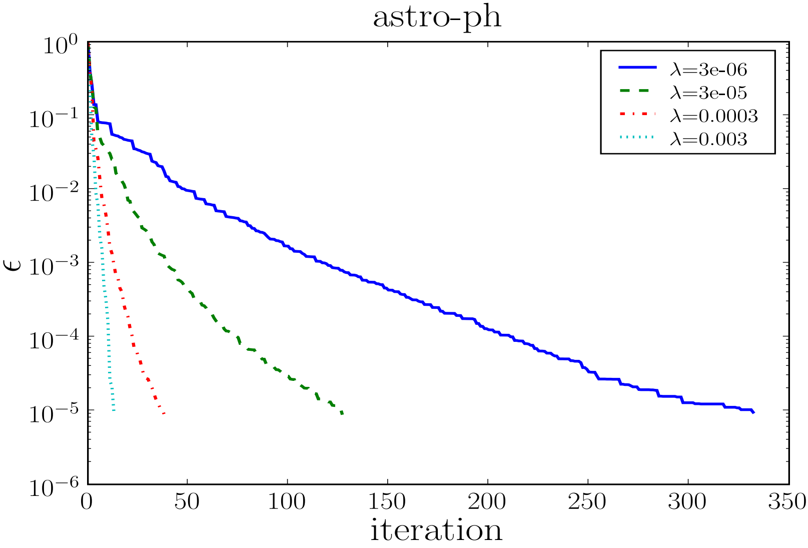
- Give lower bound on increase in dual problem in terms of γ_i and the subgradient $\partial_w [R_{\mathrm{emp}}[w] + \lambda\Omega[w]]$.
- For unbounded Hessian we have $\delta\gamma = O(\gamma^2)$.
- For bounded Hessian we have $\delta\gamma = O(\gamma)$.

Convergence

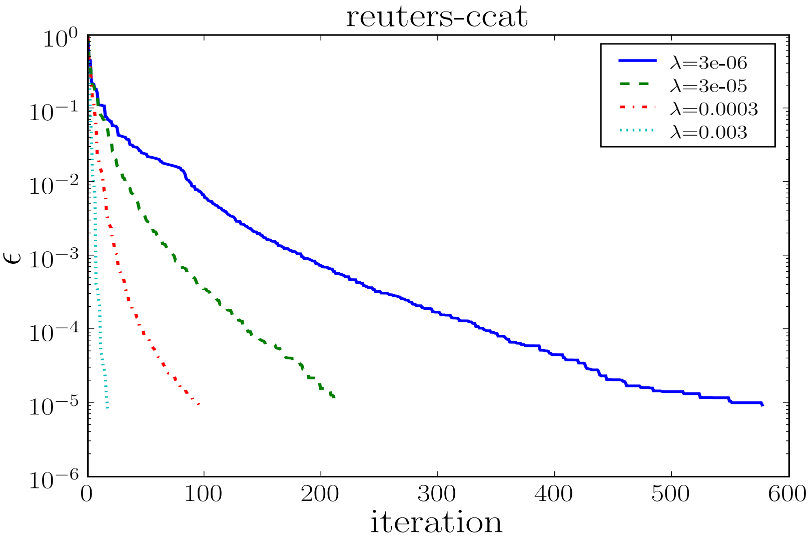
- Solve difference equation in γ_t to get desired result.

Outline

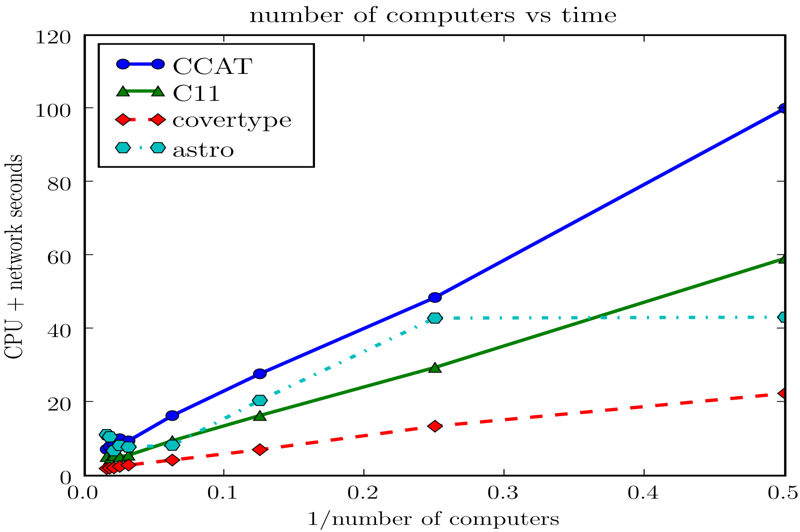
Scalability: Astrophysics dataset



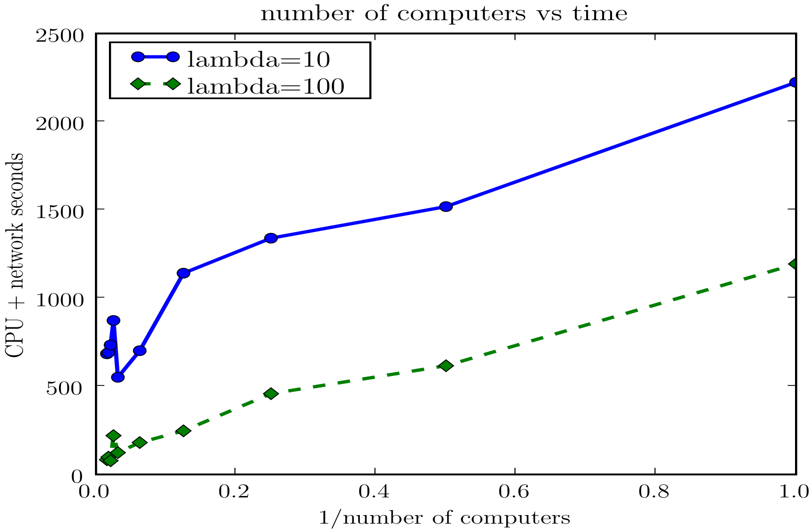
Scalability: Reuters dataset



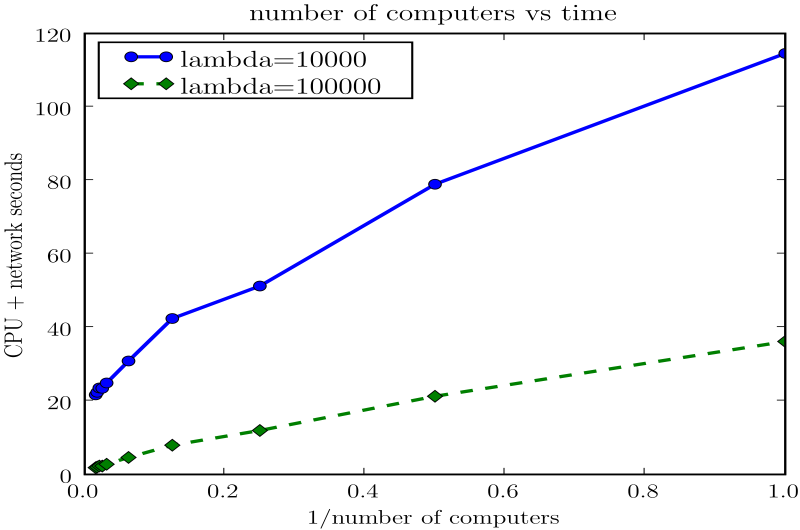
Parallelization: classification



Parallelization: ranking



Parallelization: ordinal regression



Summary

1. **Convexity in Machine Learning**
   - Linear Function Classes
   - Loss Functions
   - Regularization
2. **Algorithm**
   - Bundle Methods
   - Dual Optimization Problem
3. **Convergence**
   - Main Result
   - Proof Idea
4. **Experiments**