Outline

1 Motivation
2 Cutting Plane Methods
3 Non Smooth Functions
4 Bundle Methods
5 BMRM
6 Convergence Analysis
7 Experiments
8 Lower Bounds
9 References
Regularized Risk Minimization

Objective Function

- Training data: \( \{x_1, \ldots, x_m\} \)
- Labels: \( \{y_1, \ldots, y_m\} \)
- Learn a vector: \( w \)

\[
\text{minimize } J(w) := \underbrace{\lambda \Omega(w)}_{\text{Regularizer}} + \frac{1}{m} \sum_{i=1}^{m} l(x_i, y_i, w) \underbrace{\text{Risk } R_{\text{emp}}}_{\text{Risk }}
\]
Binary Classification

\[ y_i = +1 \]

\[ y_i = -1 \]
$y_i = +1$

$y_i = -1$
Binary Classification

\[ y_i = +1 \]

\[ y_i = -1 \]
Motivation

Binary Classification

\[ y_i = +1 \]
\[ \langle w, x \rangle + b = +1 \]
\[ \langle w, x \rangle + b = -1 \]
\[ \langle w, x_1 - x_2 \rangle = 2 \]
\[ \langle \frac{w}{\|w\|}, x_1 - x_2 \rangle = \frac{2}{\|w\|} \]

\[ y_i = -1 \]
\[ \{ x \mid \langle w, x \rangle + b = -1 \} \]
\[ \{ x \mid \langle w, x \rangle + b = 0 \} \]

\[ \{ x \mid \langle w, x \rangle + b = 1 \} \]
Motivation

Linear Support Vector Machines

Optimization Problem

$$\max_{w,b} \frac{2}{\|w\|}$$

s.t. \( y_i(\langle w, x_i \rangle + b) \geq 1 \) for all \( i \)
Linear Support Vector Machines

Optimization Problem

$$\min_{w,b} \quad \frac{1}{2} \|w\|^2$$

s.t. \( y_i (\langle w, x_i \rangle + b) \geq 1 \) for all \( i \)
Linear Support Vector Machines

Optimization Problem

\[
\begin{align*}
\min_{w,b,\xi} & \quad \frac{1}{2} \|w\|^2 \\
\text{s.t.} & \quad y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i \text{ for all } i \\
& \quad \xi_i \geq 0
\end{align*}
\]
Motivation

Linear Support Vector Machines

Optimization Problem

\[ \min_{w,b,\xi} \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \]

s.t. \[ y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \text{ for all } i \]

\[ \xi_i \geq 0 \]
Linear Support Vector Machines

Motivation

Optimization Problem

$$\min_{w, b, \xi} \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i$$

s.t. $$\xi_i \geq 1 - y_i (\langle w, x_i \rangle + b)$$ for all $$i$$

$$\xi_i \geq 0$$
Linear Support Vector Machines

Optimization Problem

\[ \min_{w, b} \frac{\lambda}{2} \| w \|^2 + \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i (\langle w, x_i \rangle + b)) \]
Motivation

Linear Support Vector Machines

Optimization Problem

\[
\min_{w, b} \lambda \frac{1}{2} \| w \|^2 + \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(\langle w, x_i \rangle + b))
\]

\[\def\arraystretch{1.5}
\begin{array}{c}
\frac{\lambda}{m} \Omega(w) \\
\frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(\langle w, x_i \rangle + b))
\end{array}
\]

\[R_{emp}(w)\]
Binary Hinge Loss

$$y(\langle w, x \rangle + b)$$
Outline

1. Motivation
2. Cutting Plane Methods
3. Non Smooth Functions
4. Bundle Methods
5. BMRM
6. Convergence Analysis
7. Experiments
8. Lower Bounds
9. References
The First Order Taylor approximation globally lower bounds the function

For any $x$ and $x'$ we have

$$f(x) \geq f(x') + \langle x - x', \nabla f(x') \rangle$$
Cutting Plane Methods
Cutting Plane Methods
Cutting Plane Methods
Cutting Plane Methods
Cutting Plane Methods
Cutting Plane Methods
Cutting Plane Methods
Cutting Plane Methods
In a Nutshell

- Cutting Plane Methods work by forming the piecewise linear lower bound

\[ J(w) \geq J^\text{CP}_t(w) := \max_{1 \leq i \leq t} \left\{ J(w_{i-1}) + \langle w - w_{i-1}, s_i \rangle \right\}. \]

where \( s_i \) denotes the gradient \( \nabla J(w_{i-1}) \).

- At iteration \( t \) the set \( \{w_i\}_{i=0}^{t-1} \) is augmented by

\[ w_t := \arg\min_w J^\text{CP}_t(w). \]

- Stop when the duality gap

\[ \epsilon_t := \min_{0 \leq i \leq t} J(w_i) - J^\text{CP}_t(w_t) \]

falls below a pre-specified threshold \( \epsilon \).
In a Nutshell

- Cutting Plane Methods work by forming the piecewise linear lower bound

\[ J(w) \geq J^\text{CP}_t(w) := \max_{1 \leq i \leq t} \left\{ J(w_{i-1}) + \langle w - w_{i-1}, s_i \rangle \right\}. \]

where \( s_i \) denotes the gradient \( \nabla J(w_{i-1}) \).

- At iteration \( t \) the set \[ \{w_i\}_{i=0}^{t-1} \] is augmented by

\[ w_t := \arg\min_w J^\text{CP}_t(w). \]

- Stop when the duality gap

\[ \epsilon_t := \min_{0 \leq i \leq t} J(w_i) - J^\text{CP}_t(w_t) \]

falls below a pre-specified threshold \( \epsilon \).
Cutting Plane Methods

In a Nutshell

- Cutting Plane Methods work by forming the piecewise linear lower bound

\[ J(w) \geq J_t^{CP}(w) := \max_{1 \leq i \leq t} \left\{ J(w_{i-1}) + \langle w - w_{i-1}, s_i \rangle \right\}. \]

where \( s_i \) denotes the gradient \( \nabla J(w_{i-1}) \).

- At iteration \( t \) the set \( \{w_i\}_{i=0}^{t-1} \) is augmented by

\[ w_t := \arg\min_w J_t^{CP}(w). \]

- Stop when the duality gap

\[ \epsilon_t := \min_{0 \leq i \leq t} J(w_i) - J_t^{CP}(w_t) \]

falls below a pre-specified threshold \( \epsilon \).
Outline

1 Motivation
2 Cutting Plane Methods
3 Non Smooth Functions
4 Bundle Methods
5 BMRM
6 Convergence Analysis
7 Experiments
8 Lower Bounds
9 References
What if the Function is NonSmooth?

The piecewise linear function

\[ J(w) := \max_i \langle u_i, w \rangle \]

is convex but not differentiable at the kinks!
A subgradient at $w'$ is any vector $s$ which satisfies

$$J(w) \geq J(w') + \langle w - w', s \rangle$$

for all $w$. 

Set of all subgradients is denoted as $\partial J(w)$. 

Non Smooth Functions

Subgradients to the Rescue

N. Vishwanathan (Purdue University)
A subgradient at $w'$ is any vector $s$ which satisfies

$$J(w) \geq J(w') + \langle w - w', s \rangle$$

for all $w$.

Set of all subgradients is denoted as $\partial J(w)$.
A subgradient at $w'$ is any vector $s$ which satisfies

$$J(w) \geq J(w') + \langle w - w', s \rangle \text{ for all } w$$

Set of all subgradients is denoted as $\partial J(w)$
Good News!

Cutting Plane Methods work with subgradients

Just choose an arbitrary one
Good News!

Cutting Plane Methods work with subgradients

Just choose an arbitrary one

Then what is the bad news?
Bad News
Outline

1 Motivation
2 Cutting Plane Methods
3 Non Smooth Functions
4 Bundle Methods
5 BMRM
6 Convergence Analysis
7 Experiments
8 Lower Bounds
9 References
Stabilized Cutting Plane Method

proximal: \( w_t := \arg\min_w \left\{ \frac{\zeta_t}{2} \| w - \hat{w}_{t-1} \|^2 + J_t^{CP}(w) \right\} \)

trust region: \( w_t := \arg\min_w \left\{ J_t^{CP}(w) \text{ s.t. } \frac{1}{2} \| w - \hat{w}_{t-1} \|^2 \leq \kappa_t \right\} \)

level set: \( w_t := \arg\min_w \left\{ \frac{1}{2} \| w - \hat{w}_{t-1} \|^2 \text{ s.t. } J_t^{CP}(w) \leq \tau_t \right\} \)

Two Kinds of Steps/Iterations

- Null Step: Enrich the local model of the objective function
- Serious Step: Decrease the objective function value
### Stabilized Cutting Plane Method

- **proximal:** 
  \[
  w_t := \arg\min_w \left\{ \frac{\zeta_t}{2} \| w - \hat{w}_{t-1} \|^2 + J_t^{CP}(w) \right\}
  \]

- **trust region:** 
  \[
  w_t := \arg\min_w \left\{ J_t^{CP}(w) \text{ s.t. } \frac{1}{2} \| w - \hat{w}_{t-1} \|^2 \leq \kappa_t \right\}
  \]

- **level set:** 
  \[
  w_t := \arg\min_w \left\{ \frac{1}{2} \| w - \hat{w}_{t-1} \|^2 \text{ s.t. } J_t^{CP}(w) \leq \tau_t \right\}
  \]

### Two Kinds of Steps/Iterations

- **Null Step:** Enrich the local model of the objective function
- **Serious Step:** Decrease the objective function value
Bundle Methods

Stabilized Cutting Plane Method

proximal: \( w_t := \arg\min_w \left\{ \frac{\xi_t}{2} \| w - \hat{w}_{t-1} \|_2^2 + J_{t}^{\text{CP}}(w) \right\} \)

trust region: \( w_t := \arg\min_w \left\{ J_{t}^{\text{CP}}(w) \text{ s.t. } \frac{1}{2} \| w - \hat{w}_{t-1} \|_2^2 \leq \kappa_t \right\} \)

level set: \( w_t := \arg\min_w \left\{ \frac{1}{2} \| w - \hat{w}_{t-1} \|_2^2 \text{ s.t. } J_{t}^{\text{CP}}(w) \leq \tau_t \right\} \)

Two Kinds of Steps/Iterations

- Null Step: Enrich the local model of the objective function
- Serious Step: Decrease the objective function value
Bundle Methods

Stabilized Cutting Plane Method

proximal: \( w_t := \arg\min_w \left\{ \frac{\xi_t}{2} \| w - \hat{w}_{t-1} \|^2 + J_{CP}^t(w) \right\} \)

trust region: \( w_t := \arg\min_w \left\{ J_{CP}^t(w) \text{ s.t. } \frac{1}{2} \| w - \hat{w}_{t-1} \|^2 \leq \kappa_t \right\} \)

level set: \( w_t := \arg\min_w \left\{ \frac{1}{2} \| w - \hat{w}_{t-1} \|^2 \text{ s.t. } J_{CP}^t(w) \leq \tau_t \right\} \)

Two Kinds of Steps/Iterations

- Null Step: Enrich the local model of the objective function
- Serious Step: Decrease the objective function value

Both involve expensive function and gradient evaluation
Outline

1. Motivation
2. Cutting Plane Methods
3. Non Smooth Functions
4. Bundle Methods
5. BMRM
6. Convergence Analysis
7. Experiments
8. Lower Bounds
9. References
Key Observation

The regularized risk already comes with stabilization built in

$$\min_w J(w) := \lambda \Omega(w) + \frac{1}{m} \sum_{i=1}^{m} l(x_i, y_i, w)$$

Bundle Method for Regularized Risk Minimization (BMRM)

1: input & initialization: $\epsilon \geq 0$, $w_0$, $t \leftarrow 0$
2: repeat
3: $t \leftarrow t + 1$
4: Compute $a_t \in \partial_w R_{\text{emp}}(w_{t-1})$ and $b_t \leftarrow R_{\text{emp}}(w_{t-1}) - \langle w_{t-1}, a_t \rangle$
5: Update model: $R_{t}^{\text{CP}}(w) := \max_{1 \leq i \leq t} \{ \langle w, a_i \rangle + b_i \}$
6: $w_t \leftarrow \arg\min_w J_t(w) := \lambda \Omega(w) + R_{t}^{\text{CP}}(w)$
7: $\epsilon_t \leftarrow \min_{0 \leq i \leq t} J(w_i) - J_t(w_t)$
8: until $\epsilon_t \leq \epsilon$
Key Observation

The regularized risk already comes with stabilization built in

\[
\min_w J(w) := \underbrace{\lambda \Omega(w)}_{\text{Regularizer}} + \frac{1}{m} \sum_{i=1}^{m} l(x_i, y_i, w) \underbrace{\text{Risk } R_{\text{emp}}}_{\text{Risk}}
\]

Bundle Method for Regularized Risk Minimization (BMRM)

1: **input & initialization:** \( \epsilon \geq 0, \ w_0, \ t \leftarrow 0 \)
2: **repeat**
3: \( \ t \leftarrow t + 1 \)
4: Compute \( a_t \in \partial_w R_{\text{emp}}(w_{t-1}) \) and \( b_t \leftarrow R_{\text{emp}}(w_{t-1}) - \langle w_{t-1}, a_t \rangle \)
5: Update model: \( R_{t}^{CP}(w) := \max_{1 \leq i \leq t} \{ \langle w, a_i \rangle + b_i \} \)
6: \( w_t \leftarrow \arg \min_w J_t(w) := \lambda \Omega(w) + R_{t}^{CP}(w) \)
7: \( \epsilon_t \leftarrow \min_{0 \leq i \leq t} J(w_i) - J_t(w_t) \)
8: **until** \( \epsilon_t \leq \epsilon \)
Key Observation

The regularized risk already comes with stabilization built in

\[
\begin{align*}
\text{minimize } & J(w) := \underbrace{\lambda \Omega(w)}_{\text{Regularizer}} + \frac{1}{m} \sum_{i=1}^{m} l(x_i, y_i, w) \\
\end{align*}
\]

Bundle Method for Regularized Risk Minimization (BMRM)

1: input & initialization: \( \epsilon \geq 0, w_0, t \leftarrow 0 \)
2: repeat
3: \( t \leftarrow t + 1 \)
4: Compute \( a_t \in \partial w R_{\text{emp}}(w_{t-1}) \) and \( b_t \leftarrow R_{\text{emp}}(w_{t-1}) - \langle w_{t-1}, a_t \rangle \)
5: Update model: \( R_{t, CP}(w) := \max_{1 \leq i \leq t} \{ \langle w, a_i \rangle + b_i \} \)
6: \( w_t \leftarrow \arg\min_w J_t(w) := \lambda \Omega(w) + R_{t, CP}(w) \)
7: \( \epsilon_t \leftarrow \min_{0 \leq i \leq t} J(w_i) - J_t(w_t) \)
8: until \( \epsilon_t \leq \epsilon \)
Outline

1 Motivation
2 Cutting Plane Methods
3 Non Smooth Functions
4 Bundle Methods
5 BMRM
6 Convergence Analysis
7 Experiments
8 Lower Bounds
9 References
Convergence Analysis

Convergence Rates

**Theorem**

Assume

- $\|\partial R_{\text{emp}}(w)\| \leq G$ for all $w$
- $\|\partial^2 \Omega(w)\| \geq H$ for all $w$

For any $\epsilon < 4G^2/H\lambda$ BMRM converges to the desired precision after

$$n \leq \log_2 \frac{H\lambda J(0)}{G^2} + \frac{8G^2}{H\lambda\epsilon} - 1$$

steps. Furthermore if the norm of the Hessian of $J(w)$ is bounded by $\bar{H}$, convergence to any $\epsilon \leq \bar{H}/2$ takes at most the following number of steps:

$$n \leq \log_2 \frac{H\lambda J(0)}{4G^2} + \frac{4}{H\lambda} \max \left[ 0, \bar{H} - \frac{8G^2}{H\lambda} \right] + \frac{4\bar{H}}{H\lambda} \log \frac{\bar{H}}{2\epsilon}$$
Proof Intuition

Let $A = [a_1, \ldots, a_t]$ and $b = [b_1, \ldots, b_t]$ where $a_t \in \partial R_{\text{emp}}(w_{t-1})$ and $b_t := R_{\text{emp}}(w_{t-1}) - \langle w_{t-1}, a_t \rangle$. The dual problem of

$$w_t = \arg\min_{w \in \mathbb{R}^d} \left\{ J_t(w) := \frac{\lambda}{2} \|w\|^2 + \max_{1 \leq i \leq t} \langle w, a_i \rangle + b_i \right\}$$

is

$$R_{\text{CP}}^t(w) \quad \text{is}$$

$$R_t \quad \text{CP} \quad (w)$$

$$\alpha_t = \arg\max_{\alpha \in \mathbb{R}^t} \left\{ -\frac{1}{2\lambda} \alpha^\top A^\top A \alpha + \alpha^\top b \right\} \text{ s.t. } \alpha \geq 0, \|\alpha\|_1 = 1.$$
Proof Intuition

Lower bound improvement in gap due to this maximization

\[ \alpha_t = \arg\max_{\alpha \in \mathbb{R}^t} \left\{ -\frac{1}{2\lambda} \alpha^\top A^\top A \alpha + \alpha^\top b \right\} \text{ s.t. } \alpha \geq 0, \|\alpha\|_1 = 1. \]

by improvement in gap due to 1-d maximization

\[
\arg\max_{\eta \in \mathbb{R}} \left\{ -\frac{1}{2\lambda} \left[ (1 - \eta)\alpha_{t-1}, \eta \right] A^\top A \left[ \begin{array}{c} (1 - \eta)\alpha_{t-1} \\ \eta \end{array} \right] + b^\top \left[ \begin{array}{c} (1 - \eta)\alpha_{t-1} \\ \eta \end{array} \right] \right\} \text{ s.t. } \eta \in [0, 1].
\]
Proof Intuition

Since function is strongly convex we can show

$$\epsilon_t - \epsilon_{t+1} \geq \frac{\epsilon_t}{2} \min(1, H\lambda\epsilon_t/4G^2).$$

Claim follows by using induction.
Comparision with Other Proofs

- Best know rates for general bundle methods is $O(1/\epsilon^3)$
- Our solver is specialized and hence better rates of convergence
- Results improve upon those of Tsochantaridis et al. who show $O(1/\epsilon^2)$ rates for a cutting plane based solver
Outline

1. Motivation
2. Cutting Plane Methods
3. Non Smooth Functions
4. Bundle Methods
5. BMRM
6. Convergence Analysis
7. Experiments
8. Lower Bounds
9. References
Convergence Behavior: Binary Classification

- RCV1: 677,399 examples, 47236 dimensions

**Iterations vs approximation gap**

![Graph showing the relationship between iterations and approximation gap for different values of λ and complexity bounds O(1/ε) and O(log(1/ε)).]
Convergence Behavior: Binary Classification

- News20: 19,954 examples, 1,355,191 dimensions

**Iterations vs approximation gap**

- $\lambda = 1e^{-3}$
- $\lambda = 1e^{-4}$
- $\lambda = 1e^{-5}$
- $\lambda = 1e^{-6}$

$O(1/\epsilon)$
$O(\log(1/\epsilon))$
Convergence Behavior: Binary Classification

- Worm: 1,026,036 examples, 804 dimensions

Iterations vs approximation gap

\[
\lambda = 1 \times 10^{-3}, \quad \lambda = 1 \times 10^{-4}, \quad \lambda = 1 \times 10^{-5}, \quad \lambda = 1 \times 10^{-6}
\]

\[
O(\frac{1}{\epsilon}), \quad O(\log(\frac{1}{\epsilon}))
\]
Outline

1. Motivation
2. Cutting Plane Methods
3. Non Smooth Functions
4. Bundle Methods
5. BMRM
6. Convergence Analysis
7. Experiments
8. Lower Bounds
9. References
Lower Bounds

Are the Rates Optimal?

Counter Example

Given $\epsilon > 0$, define $m = \frac{2}{\epsilon}$, $y_i = (-1)^i$, $x_i \in \mathbb{R}^{m+1}$ such that

$$x_i = (-1)^i \begin{bmatrix} \sqrt{m} \\ 0 \\ \vdots \\ 0 \\ m \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$
Are the Rates Optimal?

Objective Function

- Set $\lambda = 1$. Then the regularized risk is

$$J(w) = \frac{1}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i \langle x_i, w \rangle)$$

$$= \frac{1}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - \sqrt{mw_1 - mw_{i+1}}).$$

- Minimizer $w^* = \left(\frac{1}{2\sqrt{m}}, \frac{1}{2m}, \frac{1}{2m}, \ldots, \frac{1}{2m}\right)^\top$ with $J(w^*) = \frac{1}{4m}$
Are the Rates Optimal?

**Theorem**

Let $w_0 = \left( \frac{1}{\sqrt{m}}, 0, 0, \ldots \right)^\top$. Then

$$\min_{1 \leq i \leq t} J(w_i) - J(w^*) > \epsilon \text{ for all } t < \frac{2}{3\epsilon}.$$

- The crux of the proof is to show that $w_t = \left( \frac{1}{\sqrt{m}}, \frac{1}{t}, \ldots, \frac{1}{t}, 0, \ldots \right)^\top$. 

S.V. N. Vishwanathan (Purdue University)  Optimization for Machine Learning  27 / 30
Understanding the Lower Bounds?

What do the Upper Bounds Guarantee?

\[ \exists \ c, \ \forall \ \epsilon > 0, \ \forall \ J \in \mathcal{F}, \ T(\epsilon; J) \leq \frac{c}{\epsilon} \]

What do the Lower Bounds Guarantee?

\[ \forall \ \epsilon > 0, \ \exists \ c, \ \exists \ J_\epsilon \in \mathcal{F}, \ s.t. \ T(\epsilon; J_\epsilon) \geq \frac{c}{\epsilon} \]
Outline

1. Motivation
2. Cutting Plane Methods
3. Non Smooth Functions
4. Bundle Methods
5. BMRM
6. Convergence Analysis
7. Experiments
8. Lower Bounds
9. References

