# Optimization for Machine Learning

## Lecture 3: Bundle Methods

S.V. N. (vishy) Vishwanathan

Purdue University
vishy@purdue.edu

July 11, 2012

## Outline

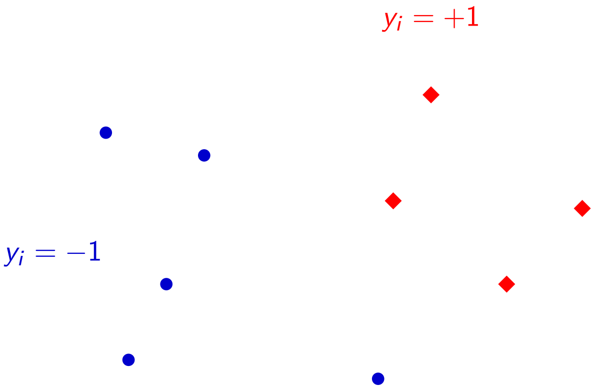
# Regularized Risk Minimization

## Objective Function

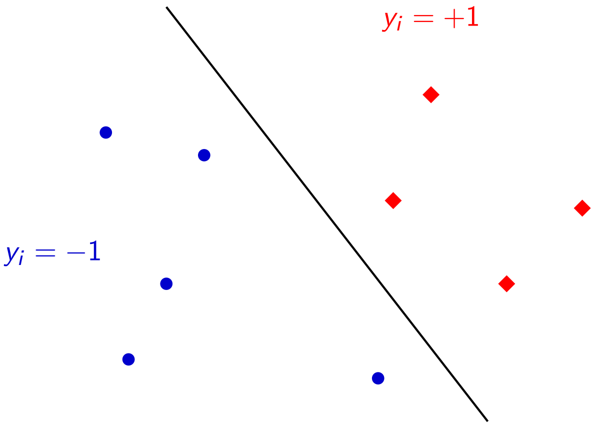
- Training data: $\{x_1, \ldots, x_m\}$
- Labels: $\{y_1, \ldots, y_m\}$
- Learn a vector: $w$

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

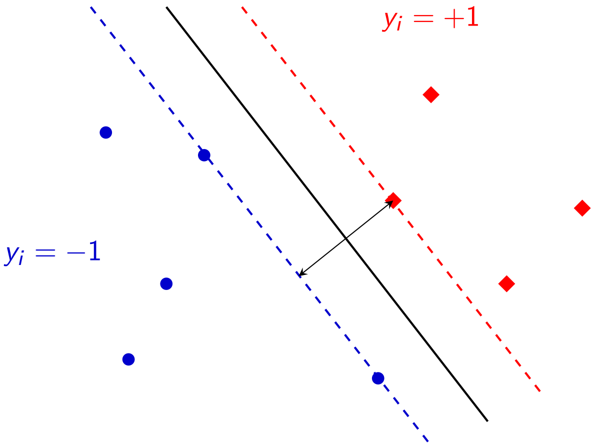
## Binary Classification



## Binary Classification

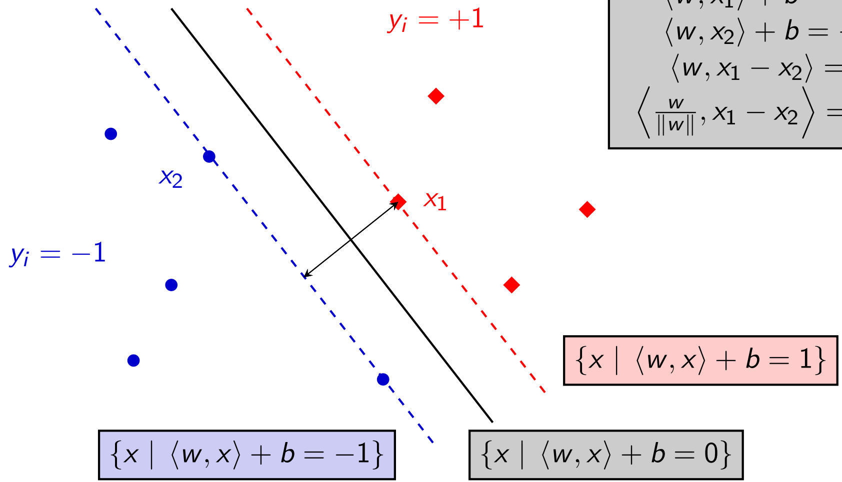


## Binary Classification



## Binary Classification

$$\langle w, x_1 \rangle + b = +1$$
$$\langle w, x_2 \rangle + b = -1$$
$$\langle w, x_1 - x_2 \rangle = 2$$
$$\left\langle \frac{w}{\|w\|}, x_1 - x_2 \right\rangle = \frac{2}{\|w\|}$$



## Linear Support Vector Machines

### Optimization Problem

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

$$\text{s.t.} \quad y_i(\langle w, x_i \rangle + b) \geq 1 \text{ for all } i$$

## Linear Support Vector Machines

### Optimization Problem

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i(\langle w, x_i \rangle + b) \geq 1 \text{ for all } i \end{aligned}$$

## Linear Support Vector Machines

### Optimization Problem

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

## Linear Support Vector Machines

### Optimization Problem

$$
\begin{aligned}
\min_{w,b,\xi} \quad & \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i \\
\text{s.t.} \quad & y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \text{ for all } i \\
& \xi_i \geq 0
\end{aligned}
$$

## Linear Support Vector Machines

### Optimization Problem

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

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

$$\xi_i \geq 0$$

# Linear Support Vector Machines

## Optimization Problem

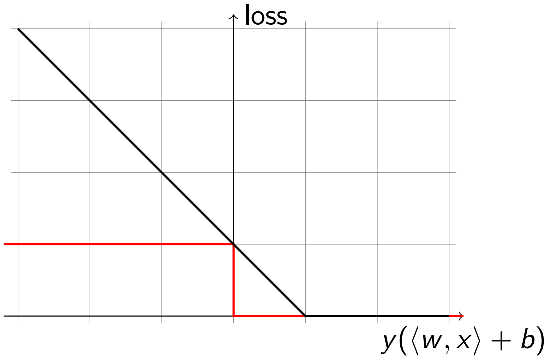
$$\min_{w,b} \quad \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max(0, 1 - y_i(\langle w, x_i \rangle + b))$$

## Linear Support Vector Machines

### Optimization Problem

$$\min_{w,b} \quad \underbrace{\frac{\lambda}{2}\|w\|^2}_{\lambda\Omega(w)} + \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max(0, 1 - y_i(\langle w, x_i\rangle + b))}_{R_{\mathrm{emp}}(w)}$$

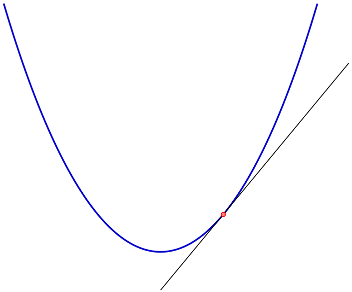
# Binary Hinge Loss



# Outline

## First Order Taylor Expansion

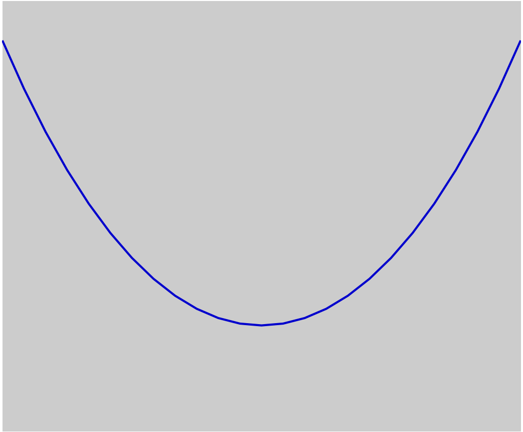
The First Order Taylor approximation globally lower bounds the function



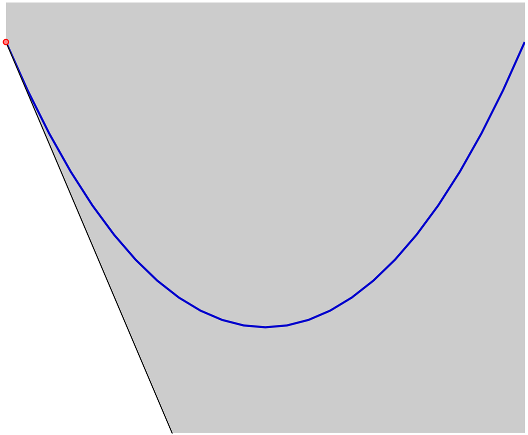
For any $x$ and $x'$ we have

$$f(x) \geq f(x') + \left\langle x - x', \nabla f(x') \right\rangle$$

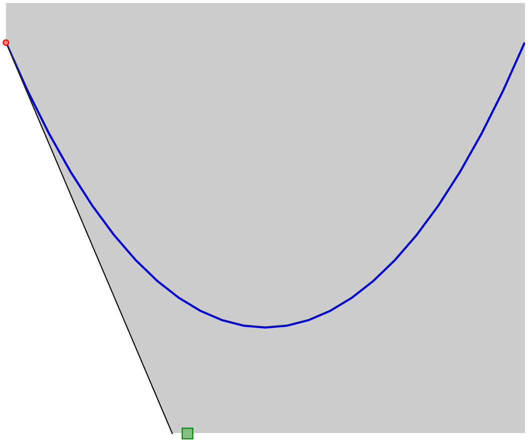
# Cutting Plane Methods



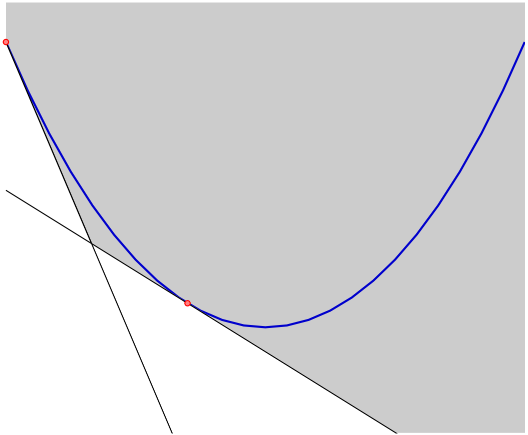
# Cutting Plane Methods



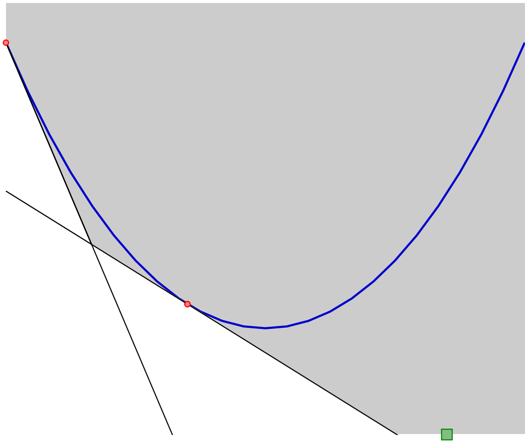
# Cutting Plane Methods



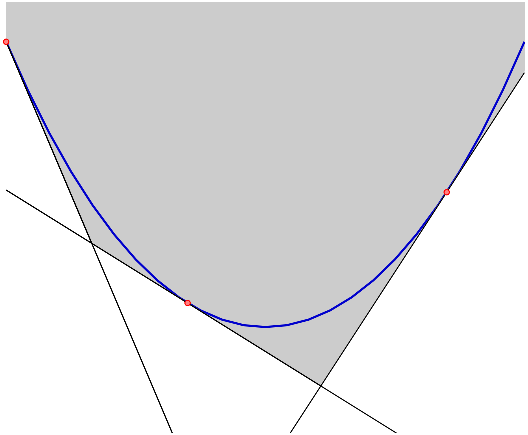
# Cutting Plane Methods



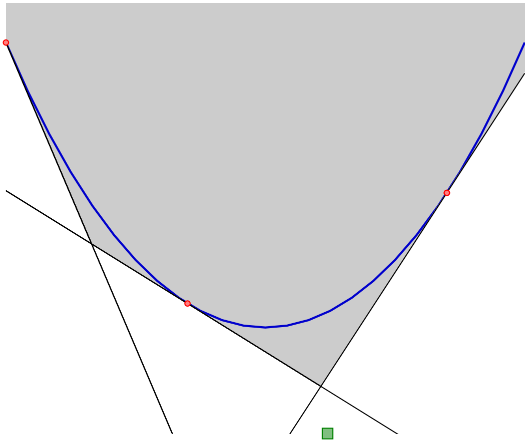
# Cutting Plane Methods



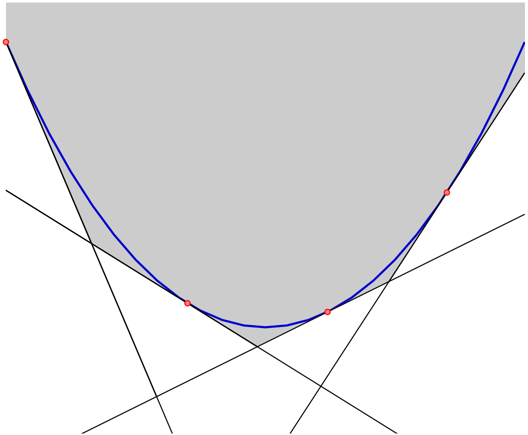
# Cutting Plane Methods



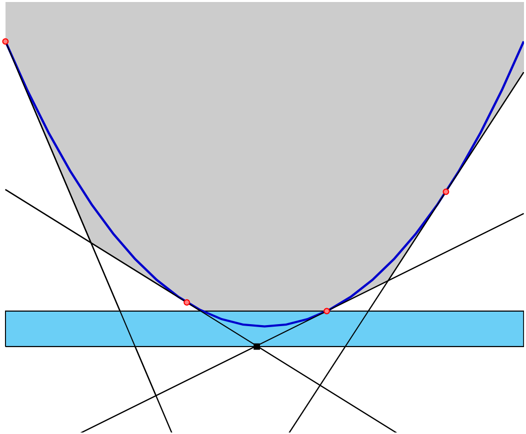
# Cutting Plane Methods



# Cutting Plane Methods



# Cutting Plane Methods



## In a Nutshell

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

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

  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

$$w_t := \operatorname*{argmin}_w J_t^{\mathrm{CP}}(w).$$

- Stop when the duality gap

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

  falls below a pre-specified threshold $\epsilon$.

## In a Nutshell

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

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

  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

$$w_t := \operatorname*{argmin}_{w} J_t^{\mathrm{CP}}(w).$$

- Stop when the duality gap

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

  falls below a pre-specified threshold $\epsilon$.

## In a Nutshell

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

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

  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

$$w_t := \operatorname*{argmin}_{w} J_t^{\mathrm{CP}}(w).$$

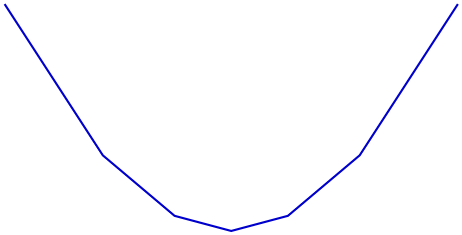
- Stop when the duality gap

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

  falls below a pre-specified threshold $\epsilon$.

## Outline

## What if the Function is NonSmooth?

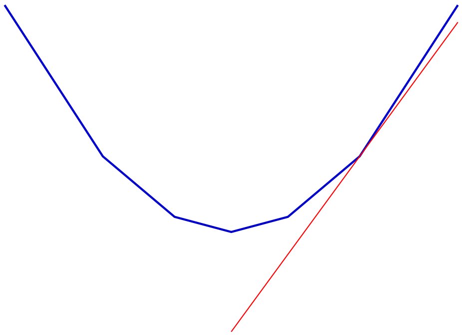


The piecewise linear function

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

is convex but not differentiable at the kinks!

## Subgradients to the Rescue

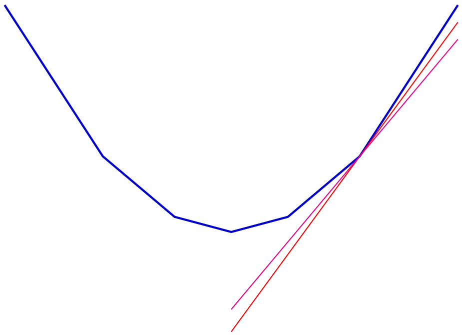


A subgradient at $w'$ is any vector $s$ which satisfies

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

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

## Subgradients to the Rescue

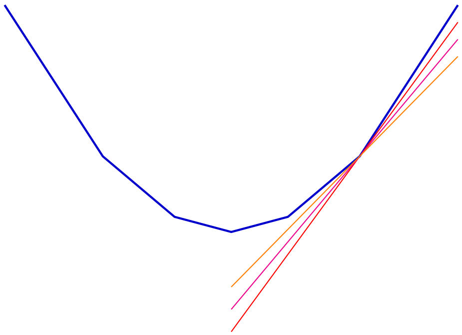


A subgradient at $w'$ is any vector $s$ which satisfies

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

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

# Subgradients to the Rescue



A subgradient at $w'$ is any vector $s$ which satisfies

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

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

## Good News!

Cutting Plane Methods work with subgradients

Just choose an arbitrary one

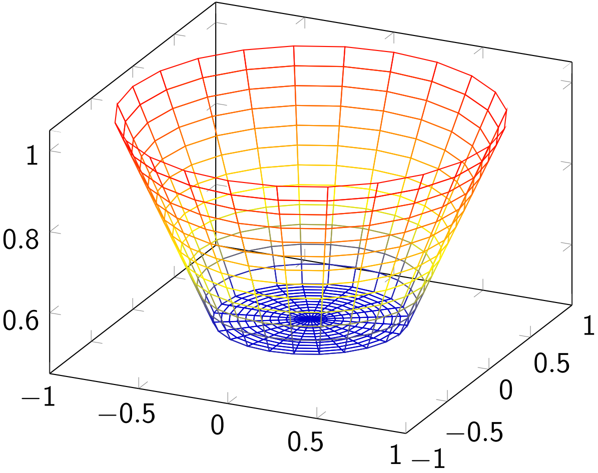
## Good News!

Cutting Plane Methods work with subgradients

Just choose an arbitrary one

Then what is the bad news?

## Bad News



# Outline

# Bundle Methods

## Stabilized Cutting Plane Method

$$\text{proximal:} \quad w_t := \operatorname*{argmin}_w \{ \tfrac{\zeta_t}{2} \|w - \hat{w}_{t-1}\|^2 + J_t^{\mathrm{CP}}(w) \}$$

$$\text{trust region:} \quad w_t := \operatorname*{argmin}_w \{ J_t^{\mathrm{CP}}(w) \text{ s.t. } \tfrac{1}{2} \|w - \hat{w}_{t-1}\|^2 \leq \kappa_t \}$$

$$\text{level set:} \quad w_t := \operatorname*{argmin}_w \{ \tfrac{1}{2} \|w - \hat{w}_{t-1}\|^2 \text{ s.t. } J_t^{\mathrm{CP}}(w) \leq \tau_t \}$$

## 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

$$\text{proximal:} \quad w_t := \operatorname*{argmin}_w \{ \tfrac{\zeta_t}{2} \| w - \hat{w}_{t-1} \|^2 + J_t^{\mathrm{CP}}(w) \}$$

$$\text{trust region:} \quad w_t := \operatorname*{argmin}_w \{ J_t^{\mathrm{CP}}(w) \text{ s.t. } \tfrac{1}{2} \| w - \hat{w}_{t-1} \|^2 \leq \kappa_t \}$$

$$\text{level set:} \quad w_t := \operatorname*{argmin}_w \{ \tfrac{1}{2} \| w - \hat{w}_{t-1} \|^2 \text{ s.t. } J_t^{\mathrm{CP}}(w) \leq \tau_t \}$$

## 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

$$\text{proximal:} \quad w_t := \operatorname*{argmin}_{w} \{ \frac{\zeta_t}{2} \|w - \hat{w}_{t-1}\|^2 + J_t^{\mathrm{CP}}(w) \}$$

$$\text{trust region:} \quad w_t := \operatorname*{argmin}_{w} \{ J_t^{\mathrm{CP}}(w) \text{ s.t. } \frac{1}{2} \|w - \hat{w}_{t-1}\|^2 \leq \kappa_t \}$$

$$\text{level set:} \quad w_t := \operatorname*{argmin}_{w} \{ \frac{1}{2} \|w - \hat{w}_{t-1}\|^2 \text{ s.t. } J_t^{\mathrm{CP}}(w) \leq \tau_t \}$$

## 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

$$\text{proximal:} \quad w_t := \operatorname*{argmin}_w \{ \frac{\zeta_t}{2} \|w - \hat{w}_{t-1}\|^2 + J_t^{\mathrm{CP}}(w) \}$$

$$\text{trust region:} \quad w_t := \operatorname*{argmin}_w \{ J_t^{\mathrm{CP}}(w) \text{ s.t. } \frac{1}{2} \|w - \hat{w}_{t-1}\|^2 \leq \kappa_t \}$$

$$\text{level set:} \quad w_t := \operatorname*{argmin}_w \{ \frac{1}{2} \|w - \hat{w}_{t-1}\|^2 \text{ s.t. } J_t^{\mathrm{CP}}(w) \leq \tau_t \}$$

## 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

## Key Observation

The regularized risk already comes with stabilization built in

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

**Bundle Method for Regularized Risk Minimization (BMRM)**

1: **input & initialization:** $\epsilon \geq 0$, $w_0$, $t \leftarrow 0$
2: **repeat**
3: $\quad t \leftarrow t + 1$
4: $\quad$ 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: $\quad$ Update model: $R_t^{\text{CP}}(w) := \max_{1 \leq i \leq t} \{\langle w, a_i \rangle + b_i\}$
6: $\quad w_t \leftarrow \text{argmin}_w \ J_t(w) := \lambda\Omega(w) + R_t^{\text{CP}}(w)$
7: $\quad \epsilon_t \leftarrow \min_{0 \leq i \leq t} J(w_i) - J_t(w_t)$
8: **until** $\epsilon_t \leq \epsilon$

## Key Observation

The regularized risk already comes with stabilization built in

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

### Bundle Method for Regularized Risk Minimization (BMRM)

1: **input & initialization:** $\epsilon \geq 0$, $w_0$, $t \leftarrow 0$
2: **repeat**
3: $\quad t \leftarrow t + 1$
4: $\quad$ 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: $\quad$ Update model: $R_t^{\text{CP}}(w) := \max_{1 \leq i \leq t}\{\langle w, a_i \rangle + b_i\}$
6: $\quad w_t \leftarrow \text{argmin}_w \, J_t(w) := \lambda\Omega(w) + R_t^{\text{CP}}(w)$
7: $\quad \epsilon_t \leftarrow \min_{0 \leq i \leq t} J(w_i) - J_t(w_t)$
8: **until** $\epsilon_t \leq \epsilon$

## Key Observation

The regularized risk already comes with stabilization built in

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

### Bundle Method for Regularized Risk Minimization (BMRM)

1: **input & initialization:** $\epsilon \geq 0$, $w_0$, $t \leftarrow 0$
2: **repeat**
3: $\quad t \leftarrow t + 1$
4: $\quad$ 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: $\quad$ Update model: $R_t^{\text{CP}}(w) := \max_{1 \leq i \leq t} \{\langle w, a_i \rangle + b_i\}$
6: $\quad w_t \leftarrow \operatorname{argmin}_w J_t(w) := \lambda\Omega(w) + R_t^{\text{CP}}(w)$
7: $\quad \epsilon_t \leftarrow \min_{0 \leq i \leq t} J(w_i) - J_t(w_t)$
8: **until** $\epsilon_t \leq \epsilon$

Look Ma no parameters!

# Outline

## Convergence Rates

### Theorem

Assume

- $\|\partial R_{\mathrm{emp}}(w)\| \leq G$ for all $w$
- $\left\|\partial^2 \Omega(w)\right\| \geq H$ 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 $a_t \in \partial R_{\mathrm{emp}}(w_{t-1})$ and $b_t := R_{\mathrm{emp}}(w_{t-1}) - \langle w_{t-1}, a_t \rangle$. The dual problem of

$$w_t = \underset{w \in \mathbb{R}^d}{\mathrm{argmin}} \{ J_t(w) := \frac{\lambda}{2} \|w\|_2^2 + \underbrace{\max_{1 \le i \le t} \langle w, a_i \rangle + b_i}_{R_t^{\mathrm{CP}}(w)} \} \quad \text{is}$$

$$\alpha_t = \underset{\alpha \in \mathbb{R}^t}{\mathrm{argmax}} \{ -\frac{1}{2\lambda} \alpha^\top A^\top A \alpha + \alpha^\top b \text{ s.t. } \alpha \ge 0, \ \|\alpha\|_1 = 1 \}.$$

## Proof Intuition

Lower bound improvement in gap due to this maximization

$$\alpha_t = \underset{\alpha \in \mathbb{R}^t}{\operatorname{argmax}} \{ -\frac{1}{2\lambda} \alpha^\top A^\top A \alpha + \alpha^\top b \text{ s.t. } \alpha \geq 0, \ \|\alpha\|_1 = 1 \}.$$

by improvement in gap due to 1-d maximization

$$\underset{\eta \in \mathbb{R}}{\operatorname{argmax}} - \frac{1}{2\lambda} \left[ \ (1-\eta)\alpha_{t-1}, \eta \ \right] A^\top A \begin{bmatrix} (1-\eta)\alpha_{t-1} \\ \eta \end{bmatrix} + b^\top \begin{bmatrix} (1-\eta)\alpha_{t-1} \\ \eta \end{bmatrix}$$

$$\text{s.t. } \eta \in [0, 1].$$

## 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

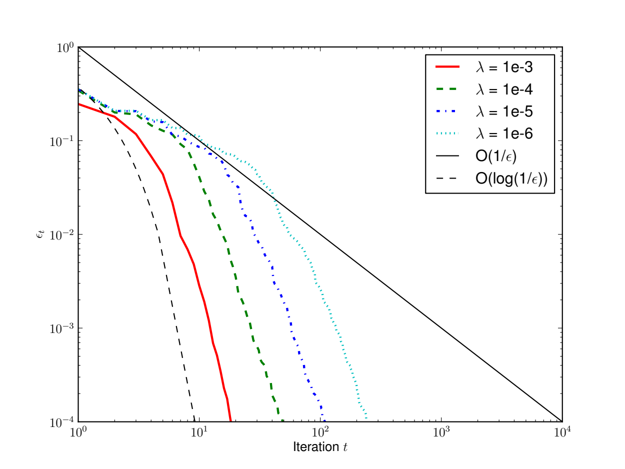
- Best know rates for general bundle methods is $O(1/\epsilon^3)$
- Our solver is specialized and hence better rates of convergence
- Results improve upon those of Tsochantaridis et al. who show $O(1/\epsilon^2)$ rates for a cutting plane based solver

# Outline

# Convergence Behavior: Binary Classification

- RCV1: 677,399 examples, 47236 dimensions

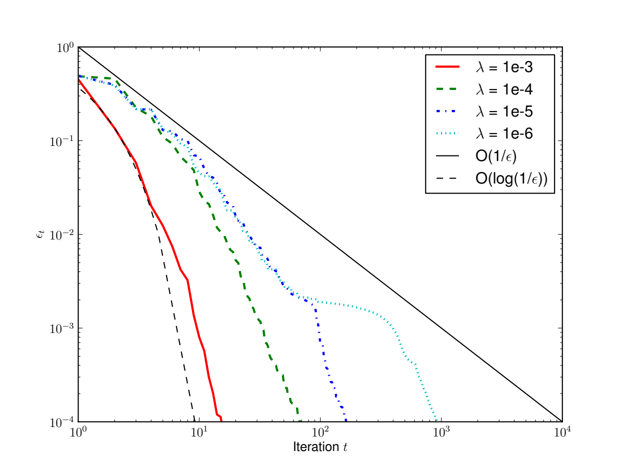
## Iterations vs approximation gap



## Convergence Behavior: Binary Classification

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

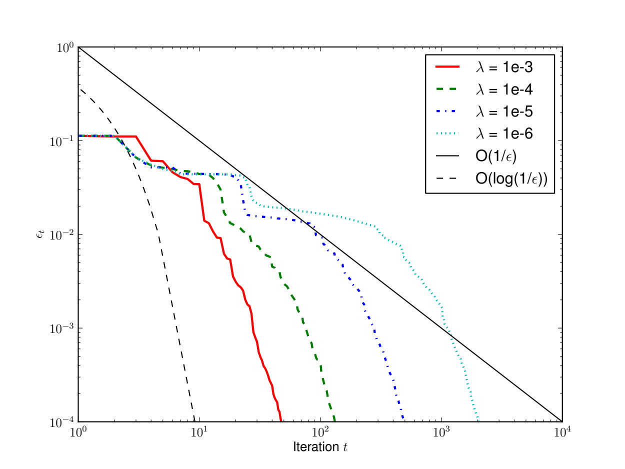
### Iterations vs approximation gap



## Convergence Behavior: Binary Classification

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

### Iterations vs approximation gap



## Outline

## 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

$$J(w) = \frac{1}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max(0, 1 - y_i \langle x_i, w\rangle)$$
$$= \frac{1}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max(0, 1 - \sqrt{m} w_1 - m w_{i+1}).$$

- 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

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

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

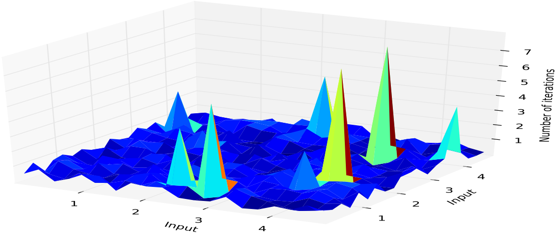
## Understanding the Lower Bounds?

### What do the Upper Bounds Guarantee?

$$\exists\ c, \forall\ \epsilon > 0, \forall\ J \in \mathfrak{F},\ \ T(\epsilon; J) \leq \frac{c}{\epsilon}$$

### What do the Lower Bounds Guarantee?

$$\forall\ \epsilon > 0, \exists\ c, \exists\ J_\epsilon \in \mathfrak{F},\ s.t.\ \ T(\epsilon; J_\epsilon) \geq \frac{c}{\epsilon}$$



## Outline

## References


- X. Zhang, A. Saha, and S. V. N. Vishwanathan. *Lower Bounds for BMRM and Faster Rates for Training SVMs*. NIPS 2010.
- C-H. Teo, S. V. N. Vishwanathan, A. Smola, and Q. V. Le. *Bundle Methods for Regularized Risk Minimization*. JMLR 11:311-365, January 2010.
- A. Smola, S. V. N. Vishwanathan, and Q. V. Le. *Bundle Methods for Machine Learning*. NIPS 2007.
- C-H. Teo, Q. Le, A. Smola, and S. V. N. Vishwanathan. *A scalable modular convex solver for regularized risk minimization*. KDD 2007.