An Online Learning Algorithm for Bilinear Models

Yuanbin Wu Shiliang Sun

East China Normal University

Introduction

- Bilinear models
- Online learning
- Regret analysis

Introduction: bilinear models

- Linear model for multi-class classification

$$h(x) = \arg\max_{y \in Y} \ w^{\intercal} \varphi(x, y)$$

- Matrix form linear model

$$h(x) = \arg\max_{y \in Y} \ \mathrm{Tr}(\ W^{\intercal} \Phi(x, y))$$

- Bilinear model

$$h(x) = \arg\max_{y \in Y} \ \alpha^{\intercal} \Phi(x, y) \beta$$

Introduction: bilinear models

- Linear model for multi-class classification

$$h(x) = \arg\max_{y \in Y} \; w^{\intercal} \varphi(x, y)$$

Matrix feature

- Matrix form linear model

$$h(x) = \arg\max_{y \in Y} \; \mathrm{Tr}(\, W^{\intercal} \Phi(x, y))$$

Rank 1 constraint on W

- Bilinear model

$$h(x) = \arg\max_{y \in Y} \; \alpha^{\intercal} \Phi(x, y) \beta$$

Introduction: online learning

- Online convex optimization
  - Convexity is violated by rank constraints
  - $\Omega_1 = \{W|\mathrm{rank}(W) \leq 1\}$ is not a convex set
- The primal dual perspective can help
  - The dual problem is always convex
- Gradients for matrix norms
  - Singular value decomposition

Introduction: regret analysis

- The regret of an online algorithm w.r.t. strategy U

$$R_N(U) = \frac{1}{N}\sum_{t=1}^{N} L_t(W_t) - \frac{1}{N}\sum_{t=1}^{N} L_t(U).$$

- Bound of the Hessian (strongly smoothness)

$$f(x+y) \leq f(x) + \nabla f(x)^\intercal y + \frac{\beta}{2}\|y\|^2$$

- Can we have similar bounds for rank constrained problems?

Outlines

Bilinear Model

Definition

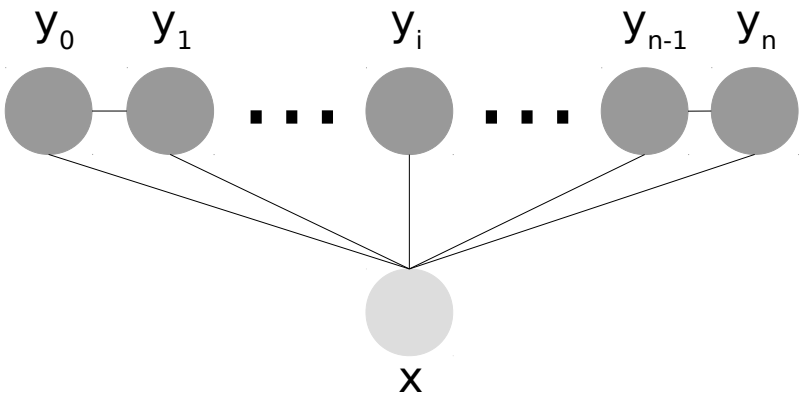
We define the bilinear model with discriminant function

$$h(x) = \arg\max_{y \in Y} \alpha^{\intercal} \Phi(x, y) \beta$$

where $\alpha \in \mathrm{R}^m, \beta \in \mathrm{R}^n$. The model parameter $W = \alpha\beta^{\intercal}$ is a rank 1 matrix.

- Why the bilinear formulation
  - semantic relations among features
  - more compact model

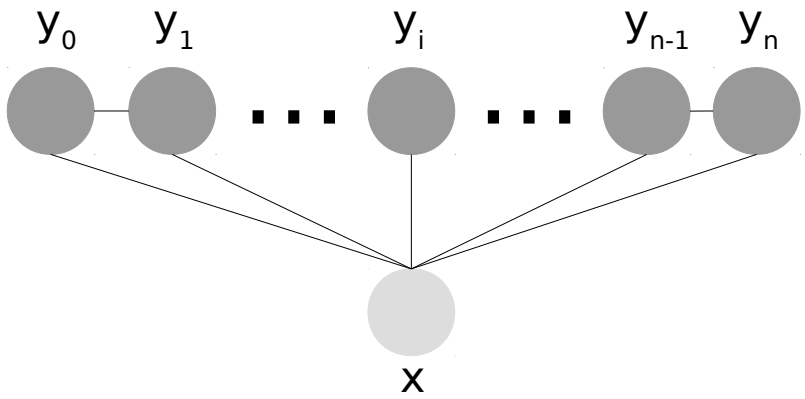
Bilinear Model



- Example: sequential labelling
  - The linear model:

$$h(x) = \arg\max_{y \in Y} \sum_{i=1}^{n} w^{\intercal} \Phi(x, y_i, y_{i-1})$$

Bilinear Model



- Example: sequential labelling
  - The linear model:

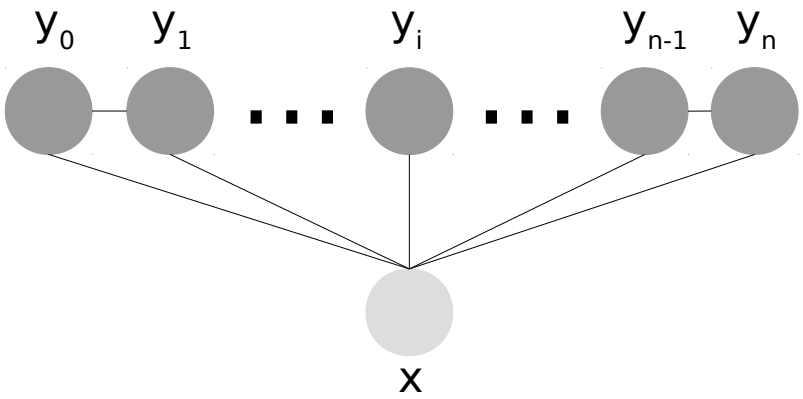
$$h(x) = \arg\max_{y \in Y} \sum_{i=1}^{n} w^{\intercal} \Phi(x, y_i, y_{i-1})$$

$$y_i y_{i-1} \quad \begin{matrix} \text{BB} & \text{BI} & \textbf{BO} & \text{IB} & \text{II} & \text{IO} & \text{OB} & \text{OI} & \text{OO} \end{matrix}$$

$$\begin{bmatrix} 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \Rightarrow \quad \begin{matrix} & \text{B} & \text{I} & \textbf{O} \\ \textbf{B} \\ \text{I} \\ \text{O} \end{matrix} \begin{bmatrix} 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{matrix} \textbf{B} \\ \text{I} \\ \text{O} \end{matrix} \begin{bmatrix} \mathbf{1} \\ 0 \\ 0 \end{bmatrix} \quad \begin{matrix} \text{B} & \text{I} & \textbf{O} \end{matrix} \begin{bmatrix} 0 & 0 & \mathbf{1} \end{bmatrix}.$$

$$\Phi(x, y_i, y_{i-1}) \qquad \Phi(x, y_i, y_{i-1}) \quad \zeta_1(x, y_i) \quad \zeta_2^{\intercal}(x, y_{i-1})$$

Bilinear Model



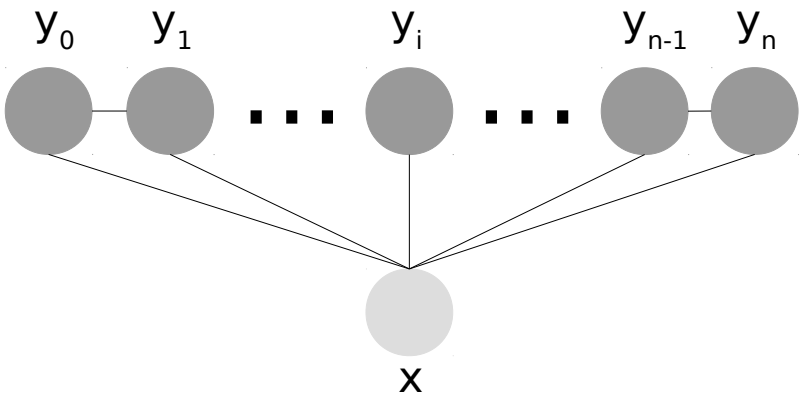
- Example: sequential labelling
  - The linear model:

$$h(x) = \arg\max_{y \in Y} \sum_{i=1}^{n} w^{\intercal} \Phi(x, y_i, y_{i-1})$$

  - The bilinear model:

$$\zeta(x, y_i) \otimes \zeta(x, y_{i-1})$$

Bilinear Model



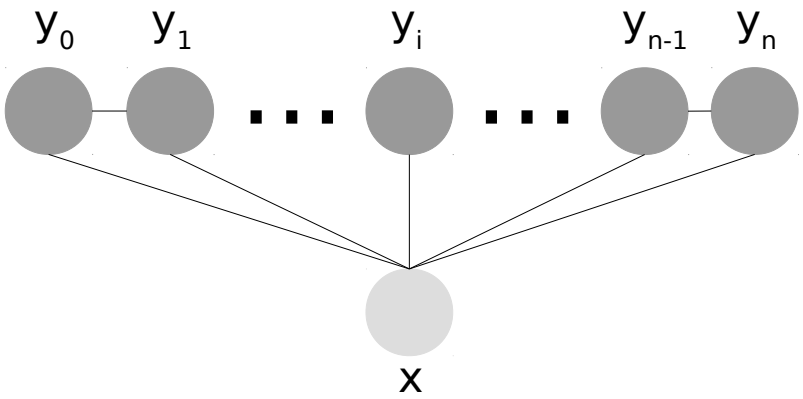
- Example: sequential labelling
  - The linear model:

$$h(x) = \arg\max_{y \in Y} \sum_{i=1}^{n} w^{\intercal} \Phi(x, y_i, y_{i-1})$$

  - The bilinear model:

$$h(x) = \arg\max_{y \in Y} \sum_{i=1}^{n} \alpha^{\intercal} \left[\zeta(x, y_i) \otimes \zeta(x, y_{i-1}) \right] \beta$$

Bilinear Model



- Example: sequential labelling
  - The linear model:

$$h(x) = \arg\max_{y \in Y} \sum_{i=1}^{n} w^{\intercal} \Phi(x, y_i, y_{i-1})$$

  - The bilinear model:

$$h(x) = \arg\max_{y \in Y} \sum_{i=1}^{n} \alpha^{\intercal} \left[\zeta(x, y_i) \otimes \zeta(x, y_{i-1}) \right] \beta$$

  - Number of parameters from $O(n^2)$ to $O(n)$

Online Learning Algorithm

- Large margin optimization problem

$$\min_{W=\alpha\beta^{\intercal}\in\Omega_1} \quad \frac{1}{2}\|W\|_{\mathrm{F}}^2 + C\sum_{j=1}^{N}[1-\langle W, \Delta\Phi^j\rangle]_+,$$

where $\quad \Delta\Phi^j \triangleq \Phi(x^j, y^j) - \Phi(x^j, h(x^j)),$

Ω_1 is the set of rank 1 matrices.

Online Learning Algorithm

- Large margin optimization problem

$$\min_{W=\alpha\beta^\intercal\in\Omega_1} \quad \frac{1}{2}\|W\|_{\mathrm{F}}^2 + C\sum_{j=1}^{N}[1-\langle W, \Delta\Phi^j\rangle]_+,$$

where $\quad \Delta\Phi^j \triangleq \Phi(x^j, y^j) - \Phi(x^j, h(x^j)),$

Ω_1 is the set of rank 1 matrices.

- Biconvex problem

$$\min_{\alpha,\beta} \quad \frac{1}{2}\|\alpha\|^2 + \frac{1}{2}\|\beta\|^2 + C\sum_{j=1}^{N}[1-\alpha^\intercal\Delta\Phi^j\beta]_+,$$

  - blockwise coordinate descent
  - degenerated cases: only solve a 0-order model on $\zeta(x, y_i)$

Online Learning Algorithm

- Our plan:
  - from the dual
  - mirror descent style updates

$$
\begin{array}{ccc}
W_{t-1} & \xrightarrow{\nabla F} & \Theta_{t-1} \\
 & & \Big\downarrow -\eta_t \nabla L_t \\
W_t & \xleftarrow{\nabla F^*} & \Theta_t
\end{array}
$$

Online Learning Algorithm

- Define $F_1(W) = \frac{1}{2}\|W\|_{\mathrm{F}}^2$ if $W \in \Omega_1$, $+\infty$ otherwise.

Online Learning Algorithm

- Define $F_1(W) = \frac{1}{2}\|W\|_{\mathrm{F}}^2$ if $W \in \Omega_1$, $+\infty$ otherwise.
- The dual problem

$$\begin{aligned}\mathcal{D}(\eta) &= \sum_{j=1}^{N} \eta_j - \max_{W \in \Omega_1} \left(\langle W, \sum_{j=1}^{N} \eta_j \Delta\Phi^j \rangle - \frac{1}{2}\|W\|_{\mathrm{F}}^2 \right) \\ &= \sum_{j=1}^{N} \eta_j - F_1^*(\Theta_N), \qquad \eta_j \in [0, C].\end{aligned}$$

Online Learning Algorithm

- Define $F_1(W) = \frac{1}{2}\|W\|_{\mathrm{F}}^2$ if $W \in \Omega_1$, $+\infty$ otherwise.
- The dual problem

$$\begin{aligned}\mathcal{D}(\eta) &= \sum_{j=1}^{N} \eta_j - \max_{W \in \Omega_1} \left(\langle W, \sum_{j=1}^{N} \eta_j \Delta\Phi^j \rangle - \frac{1}{2}\|W\|_{\mathrm{F}}^2 \right) \\ &= \sum_{j=1}^{N} \eta_j - F_1^*(\Theta_N), \qquad \eta_j \in [0, C].\end{aligned}$$

where

$$\Theta_N = \Theta_{N-1} + \eta_N \Delta\Phi^N \text{ (gradients of hinge loss, mirror space)}$$

$$F_1^*(\Theta) = \max_{W \in \Omega_1} \langle W, \Theta \rangle - \frac{1}{2}\|W\|_{\mathrm{F}}^2 \text{ (the Frenchel dual)}$$

Online Learning Algorithm

- The dual problem $\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - F_1^*(\Theta_N)$

$$\Theta_N = \Theta_{N-1} + \eta_N \Delta\Phi^N$$

$$F_1^*(\Theta) = \max_{W \in \Omega_1} \langle W, \Theta \rangle - \tfrac{1}{2} \| W \|_{\mathrm{F}}^2$$

Online Learning Algorithm

$$\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - \frac{1}{2}\|\Theta_N\|_2^2$$

- The dual problem $\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - F_1^*(\Theta_N)$

$$\Theta_N = \Theta_{N-1} + \eta_N \Delta\Phi^N$$

$$F_1^*(\Theta) = \max_{W \in \Omega_1} \langle W, \Theta \rangle - \frac{1}{2}\| W\|_{\mathrm{F}}^2$$

- Proposition: $F_1^*(\Theta) = \frac{1}{2}\|\Theta\|_2^2 = \frac{1}{2}\|\Theta\|_{\mathbf{s}(\infty)}^2 = \frac{1}{2}\sigma_1(\Theta)^2$
  - SVD has property of "the best low rank approximation"

Online Learning Algorithm

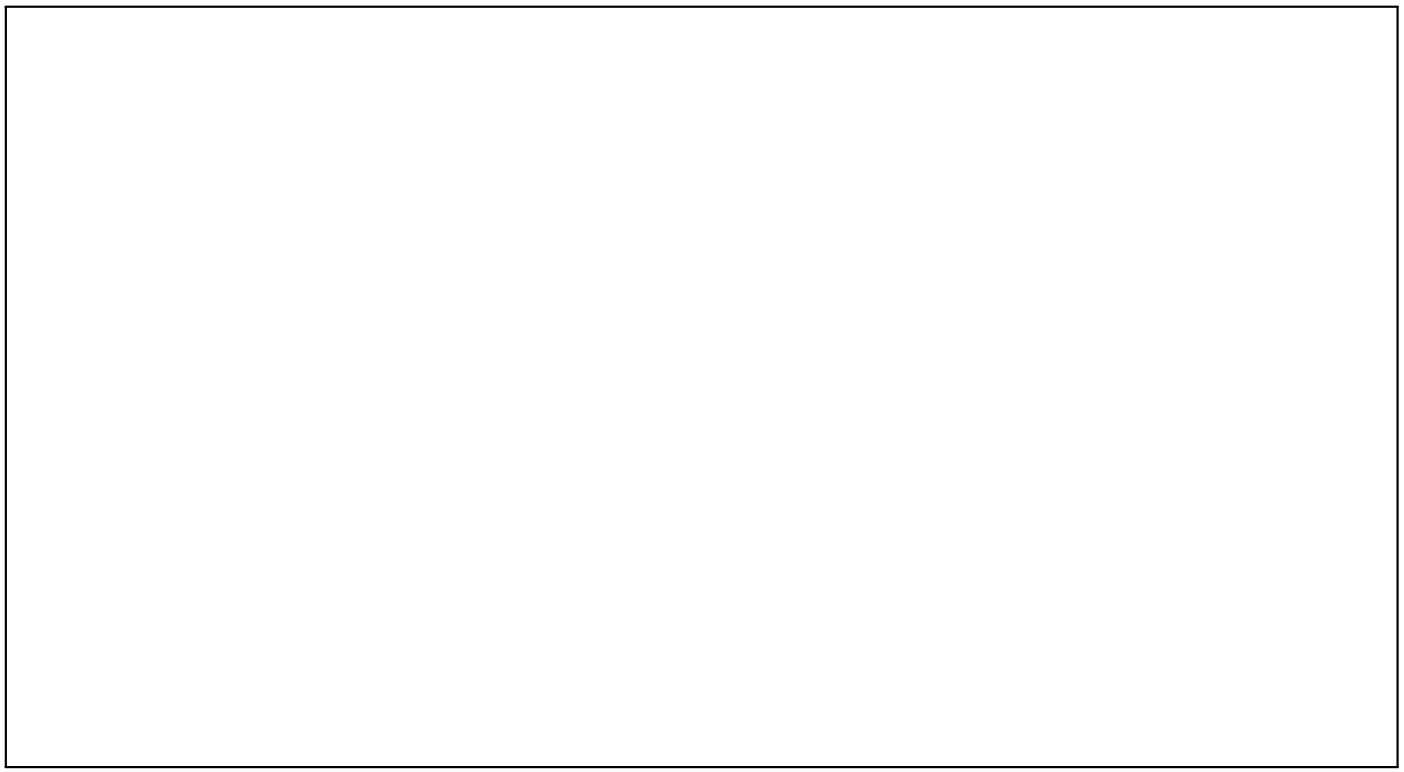
$$\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - \frac{1}{2}\|\Theta_N\|_2^2$$

- The dual problem $\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - F_1^*(\Theta_N)$

$$\Theta_N = \Theta_{N-1} + \eta_N \Delta\Phi^N$$

$$F_1^*(\Theta) = \max_{W \in \Omega_1} \langle W, \Theta \rangle - \frac{1}{2}\| W\|_{\mathrm{F}}^2$$

- A series of dual problems $\mathcal{D}_{t+1}(\eta) = \sum_{j=1}^{t} \eta_j - F_1^*(\Theta_t), t = 1, 2, \ldots, N$



Online Learning Algorithm

$$\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - \frac{1}{2}\|\Theta_N\|_2^2$$

- The dual problem $\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - F_1^*(\Theta_N)$

$$\Theta_N = \Theta_{N-1} + \eta_N \Delta\Phi^N$$

$$F_1^*(\Theta) = \max_{W \in \Omega_1} \langle W, \Theta \rangle - \frac{1}{2}\| W\|_{\mathrm{F}}^2$$

- A series of dual problems $\mathcal{D}_{t+1}(\eta) = \sum_{j=1}^{t} \eta_j - F_1^*(\Theta_t), t = 1, 2, \ldots, N$

  - uses $W_{t-1} = \alpha_{t-1}\beta_{t-1}^{\mathsf{T}}$ to predict x^t, $\bar{y}^t = h(x^t)$;

Online Learning Algorithm

$$\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - \frac{1}{2}\|\Theta_N\|_2^2$$

- The dual problem $\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - F_1^*(\Theta_N)$

$$\Theta_N = \Theta_{N-1} + \eta_N \Delta\Phi^N$$

$$F_1^*(\Theta) = \max_{W \in \Omega_1} \langle W, \Theta \rangle - \frac{1}{2}\| W\|_{\mathrm{F}}^2$$

- A series of dual problems $\mathcal{D}_{t+1}(\eta) = \sum_{j=1}^{t} \eta_j - F_1^*(\Theta_t), t = 1, 2, \ldots, N$

  - uses $W_{t-1} = \alpha_{t-1}\beta_{t-1}^{\mathsf{T}}$ to predict x^t, $\bar{y}^t = h(x^t)$;
  - sets the dual variable η_t as

$$\eta_t = \begin{cases} 0 & \bar{y}^t = y^t \\ C & \bar{y}^t \neq y^t \end{cases}$$

Online Learning Algorithm

$$\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - \frac{1}{2}\|\Theta_N\|_2^2$$

- The dual problem $\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - F_1^*(\Theta_N)$

$$\Theta_N = \Theta_{N-1} + \eta_N \Delta\Phi^N$$

$$F_1^*(\Theta) = \max_{W \in \Omega_1} \langle W, \Theta \rangle - \frac{1}{2}\| W\|_{\mathrm{F}}^2$$

- A series of dual problems $\mathcal{D}_{t+1}(\eta) = \sum_{j=1}^{t} \eta_j - F_1^*(\Theta_t), t = 1, 2, \ldots, N$

  - uses $W_{t-1} = \alpha_{t-1}\beta_{t-1}^{\mathsf{T}}$ to predict x^t, $\bar{y}^t = h(x^t)$;
  - sets the dual variable η_t as

$$\eta_t = \begin{cases} 0 & \bar{y}^t = y^t \\ C & \bar{y}^t \neq y^t \end{cases}$$

  - updates W_t:

$$W_t = \nabla F_1^*(\Theta_t) = \arg \max_{W \in \Omega_1} \langle W, \Theta_t \rangle - \frac{1}{2}\| W\|_{\mathrm{F}}^2$$

Online Learning Algorithm

$$\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - \frac{1}{2}\|\Theta_N\|_2^2$$

- The dual problem $\mathcal{D}(\eta) = \sum_{j=1}^{N} \eta_j - F_1^*(\Theta_N)$

$$\Theta_N = \Theta_{N-1} + \eta_N \Delta\Phi^N$$

$$F_1^*(\Theta) = \max_{W \in \Omega_1} \langle W, \Theta \rangle - \frac{1}{2}\| W\|_{\mathrm{F}}^2$$

- A series of dual problems $\mathcal{D}_{t+1}(\eta) = \sum_{j=1}^{t} \eta_j - F_1^*(\Theta_t), t = 1, 2, \ldots, N$

  - uses $W_{t-1} = \alpha_{t-1}\beta_{t-1}^{\mathsf{T}}$ to predict x^t, $\bar{y}^t = h(x^t)$;
  - sets the dual variable η_t as

$$\eta_t = \begin{cases} 0 & \bar{y}^t = y^t \\ C & \bar{y}^t \neq y^t \end{cases}$$

  - updates W_t:

$$W_t = \nabla F_1^*(\Theta_t) = \arg \max_{W \in \Omega_1} \langle W, \Theta_t \rangle - \frac{1}{2}\| W\|_{\mathrm{F}}^2$$

$$\overset{\sigma_1 \neq \sigma_2}{=} \sigma_1 u_1 v_1^{\mathsf{T}}$$

Online Learning Algorithm

- uses $W_{t-1} = \alpha_{t-1}\beta_{t-1}^{\mathsf{T}}$ to predict x^t,

$$\bar{y}^t = h(x^t) = \arg\max_{y \in Y} \alpha_{t-1}^{\mathsf{T}} \Delta\Phi_t(x^t, y)\beta_{t-1}$$

- sets the dual variable η_t as

$$\eta_t = \begin{cases} 0 & \bar{y}^t = y^t \\ C & \bar{y}^t \neq y^t \end{cases}$$

- updates W_t:

$$\Theta_t = \Theta_{t-1} + \eta_t \Delta\Phi^t = \sum_{i=1}^{p} \sigma_i u_i v_i^{\mathsf{T}}$$

$$W_t = \sigma_1 u_1 v_1^{\mathsf{T}}$$

Online Learning Algorithm

$$W_t = \nabla F_1^*(\Theta_t) = \sigma_1 u_1 v_1^\intercal$$

- Full SVD is expensive, only needs the leading singular vectors

Online Learning Algorithm

$$W_t = \nabla F_1^*(\Theta_t) = \sigma_1 u_1 v_1^\intercal$$

- Full SVD is expensive, only needs the leading singular vectors
- Power iteration
  - if $\sigma_1(\Theta) \neq \sigma_2(\Theta)$

$$\alpha^{(\tau+1)} = \Theta^\intercal \Theta \alpha^{(\tau)}, \qquad \frac{\alpha^{(\tau+1)}}{\|\alpha^{(\tau+1)}\|} \to u_1$$

$$\beta^{(\tau+1)} = \Theta \Theta^\intercal \beta^{(\tau)}, \qquad \frac{\beta^{(\tau+1)}}{\|\beta^{(\tau+1)}\|} \to v_1$$

Online Learning Algorithm

$$W_t = \nabla F_1^*(\Theta_t) = \sigma_1 u_1 v_1^\intercal$$

- Full SVD is expensive, only needs the leading singular vectors
- Power iteration
  - if $\sigma_1(\Theta) \neq \sigma_2(\Theta)$

$$\alpha^{(\tau+1)} = \Theta^\intercal \Theta \alpha^{(\tau)}, \qquad \frac{\alpha^{(\tau+1)}}{\|\alpha^{(\tau+1)}\|} \to u_1$$

$$\beta^{(\tau+1)} = \Theta \Theta^\intercal \beta^{(\tau)}, \qquad \frac{\beta^{(\tau+1)}}{\|\beta^{(\tau+1)}\|} \to v_1$$

  - Initial value and normalization
    - ★ $\Theta_t = \Theta_{t-1} + \eta_t \Delta\Phi^t$
    - ★ if $\Delta\Phi^t$ is "small", α_t is close to α_{t-1}
    - ★ if $\Delta\Phi^t$ is "sparse", normalization could be efficient

Regret Analysis

- The regret w.r.t. strategy U

$$R_N(U) = \frac{1}{N}\sum_{t=1}^{N} L_t(W_t) - \frac{1}{N}\sum_{t=1}^{N} L_t(U).$$

- W_t are weights at each round
- L_t is the hinge loss

Regret Analysis

- Previous analysis (mirror descent)

$$
\begin{array}{ccc}
W_{t-1} & \xrightarrow{\nabla F} & \Theta_{t-1} \\
 & & \downarrow -\eta_t \nabla L_t \\
W_t & \xleftarrow{\nabla F^*} & \Theta_t
\end{array}
$$

  - If L_t is convex and F is strongly convex, then $R_N(U) = O(\frac{1}{\sqrt{N}})$

Regret Analysis

$$F_1(W) = \frac{1}{2}\|W\|_{\mathrm{F}}^2,\ W \in \Omega_1$$

- Previous analysis (mirror descent)

$$\begin{array}{ccc} W_{t-1} & \xrightarrow{\nabla F} & \Theta_{t-1} \\ & & \downarrow -\eta_t \nabla L_t \\ W_t & \xleftarrow{\nabla F^*} & \Theta_t \end{array}$$

  - If L_t is convex and F is strongly convex, then $R_N(U) = O(\frac{1}{\sqrt{N}})$
- In bilinear model
  - $F_1(W) = \frac{1}{2}\|W\|_{\mathrm{F}}^2$ if $W \in \Omega_1$, $+\infty$ otherwise.
  - not convex
  - $F_1^{**}(W) = \frac{1}{2}\|W\|_2^2 \neq F_1$
  - The analysis of mirror descent is not directly applicable

Regret Analysis

- Lower bound of dual objective + weak duality
- Bound the increase of the dual objective

$$\begin{aligned}\Delta_t &= \mathcal{D}_{t+1}(\eta_1, \ldots, \eta_t) - \mathcal{D}_t(\eta_1, \ldots, \eta_{t-1}) \\ &= C - \frac{1}{2}\|\Theta_{t-1} + C\Delta\Phi^t\|_2^2 + \frac{1}{2}\|\Theta_{t-1}\|_2^2.\end{aligned}$$

- By the Taylor expansion:

$$\frac{1}{2}\|\Theta + E\|_2^2 \leq \frac{1}{2}\|\Theta\|_2^2 + \langle \nabla\|\Theta\|_2, E\rangle + \mathrm{vec}(E)^\intercal H(\hat{\Theta})\mathrm{vec}(E)$$
$$\text{where } \hat{\Theta} = \Theta + \theta E, \theta \in (0, 1)$$

- Bound the Hessian term

Regret Analysis

- Our result (by bounding the Hessian)
  - If $\sigma_1(\Theta) \neq \sigma_2(\Theta) > 0$,

$$\frac{1}{2}\|\Theta + E\|_2^2 \leq \frac{1}{2}\|\Theta\|_2^2 + \langle \nabla \|\Theta\|_2, E \rangle + \|E\|_{\mathrm{F}}^2 \frac{2l}{1 - \frac{\hat{\sigma}_2}{\hat{\sigma}_1}}$$

where $[\hat{\sigma}_1, \ldots, \hat{\sigma}_l] = \sigma(\hat{\Theta}), \hat{\Theta} = \Theta + \theta E, \theta \in (0, 1)$

Regret Analysis

- Our result (by bounding the Hessian)
  - If $\sigma_1(\Theta) \neq \sigma_2(\Theta) > 0$,

$$\frac{1}{2}\|\Theta + E\|_2^2 \leq \frac{1}{2}\|\Theta\|_2^2 + \langle \nabla \|\Theta\|_2, E \rangle + \|E\|_{\mathrm{F}}^2 \frac{2l}{1 - \frac{\hat{\sigma}_2}{\hat{\sigma}_1}}$$

$$\text{where } [\hat{\sigma}_1, \ldots, \hat{\sigma}_l] = \sigma(\hat{\Theta}), \hat{\Theta} = \Theta + \theta E, \theta \in (0, 1)$$

- Known result on Schatten norm (Ball et al., 1994; Kakade et al., 2012):
  - Schatten norm: $\|\Theta\|_{\mathbf{s}(p)} = \|\sigma(\Theta)\|_p$, $\|\Theta\|_{\mathbf{s}(\infty)} = \|\Theta\|_2 = \sigma_1(\Theta)$

Regret Analysis

- Our result (by bounding the Hessian)
  - If $\sigma_1(\Theta) \neq \sigma_2(\Theta) > 0$,

$$\frac{1}{2}\|\Theta + E\|_2^2 \leq \frac{1}{2}\|\Theta\|_2^2 + \langle \nabla\|\Theta\|_2, E\rangle + \|E\|_{\mathrm{F}}^2 \frac{2l}{1 - \frac{\hat{\sigma}_2}{\hat{\sigma}_1}}$$

$$\text{where } [\hat{\sigma}_1, \ldots, \hat{\sigma}_l] = \sigma(\hat{\Theta}), \hat{\Theta} = \Theta + \theta E, \theta \in (0, 1)$$

- Known result on Schatten norm (Ball et al., 1994; Kakade et al., 2012):
  - Schatten norm: $\|\Theta\|_{\mathbf{s}(p)} = \|\sigma(\Theta)\|_p$, $\|\Theta\|_{\mathbf{s}(\infty)} = \|\Theta\|_2 = \sigma_1(\Theta)$
  - for $p \in [2, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$,

$$\frac{1}{2}\|\Theta + E\|_{\mathbf{s}(p)}^2 \leq \frac{1}{2}\|\Theta\|_{\mathbf{s}(p)}^2 + \langle \nabla\|\Theta\|_{\mathbf{s}(p)}, E\rangle + \frac{\|E\|_{\mathbf{s}(q)}^2}{2(q-1)}.$$

  - The bound is trivial if $p = \infty$.

Regret Analysis

Proposition (Regret)

Assume for all $\Theta = \Theta_{t-1}, E = C\Delta\Phi^t$, the bound of Hessian holds. Then

$$R_N(U) \leq \frac{1}{2CN}\|U\|_{\mathrm{F}}^2 + \frac{2lC}{N}\sum_{t=1}^{N}\frac{\|\Delta\Phi^t\|_{\mathrm{F}}^2}{1-\frac{\hat{\sigma}_2^t}{\hat{\sigma}_1^t}}.$$

- The role of $\frac{\sigma_1^t}{\sigma_2^t}$
  - the speed of power iteration
  - the regret bound

Regret Analysis

- Bound $\frac{\sigma_1}{\sigma_2}$: margin requirement + "$\sigma_1$ is uniformly greater than σ_2"

Proposition

Assume that $\sup_{j,W} \|\Delta\Phi^j\|_2 \leq M_1$, $\sup_{j,W} \|\Delta\Phi^j\|_{\mathbf{k}(2)} \leq M_2$. *If* $M_1 > \frac{M_2}{2}$ *and* $\exists \tilde{W}$ *has margin* γ *w.r.t.* $\|\cdot\|_{\mathbf{s}(1)}$, *where* $\gamma \in (\frac{M_2}{2}, M_1)$, *then*

$$\frac{\hat{\sigma}_2^t}{\hat{\sigma}_1^t} \leq \frac{M_2 - \gamma}{\gamma}.$$

Corollary

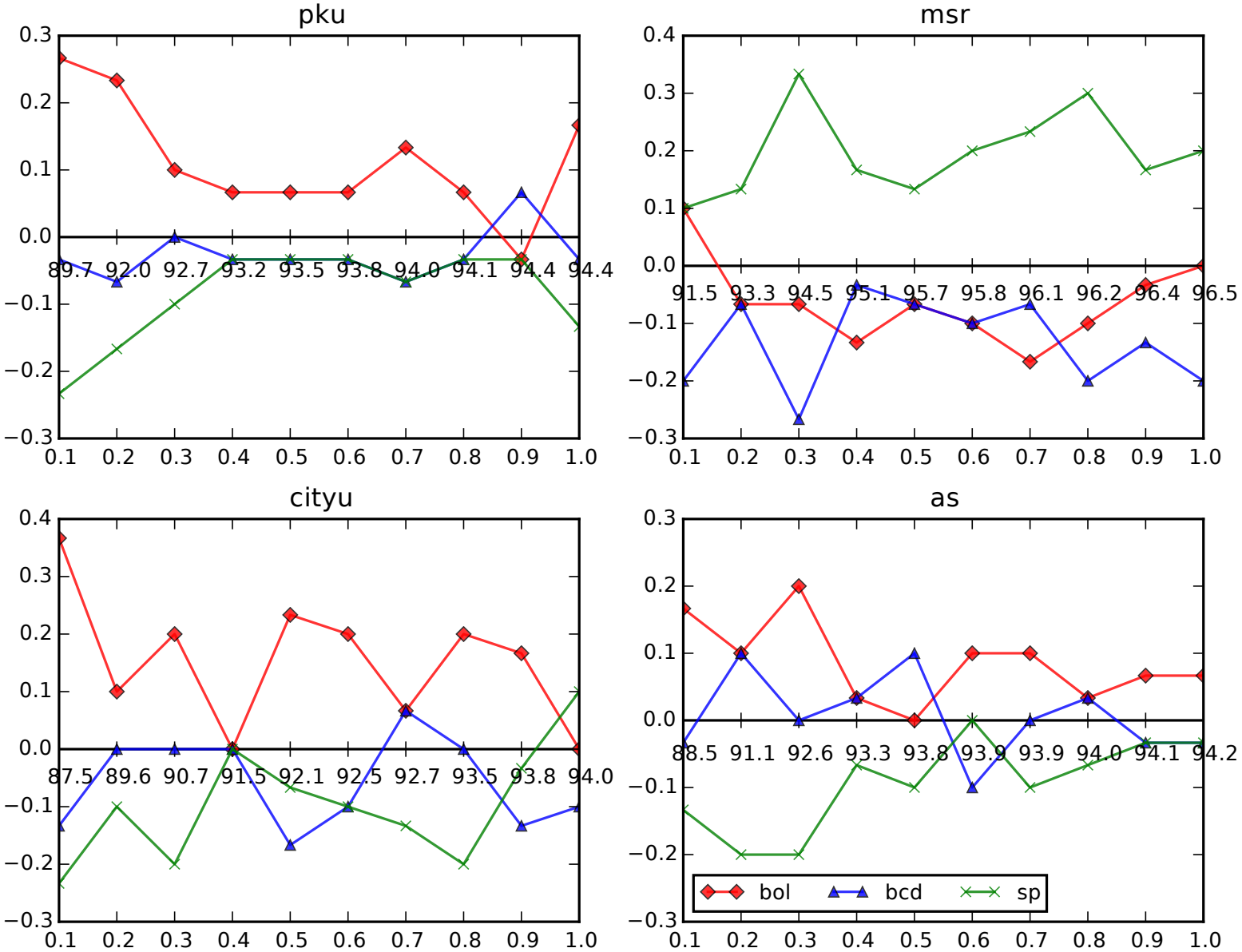
The regret is bounded by

$$R_N(U) \leq \frac{1}{2CN}\|U\|_{\mathrm{F}}^2 + 2Cl^2 M_1^2 \frac{\gamma}{2\gamma - M_2}.$$

Experiments

- Two sequential labelling tasks
  - Chinese words segmentation
  - Text chunking
- Baselines
  - Linear model (structured perceptron)
  - Blockwise coordinate descent of the biconvex problem
  - Batch learner (CRF+L_2, CRF+L_1)

Experiments



Experiments

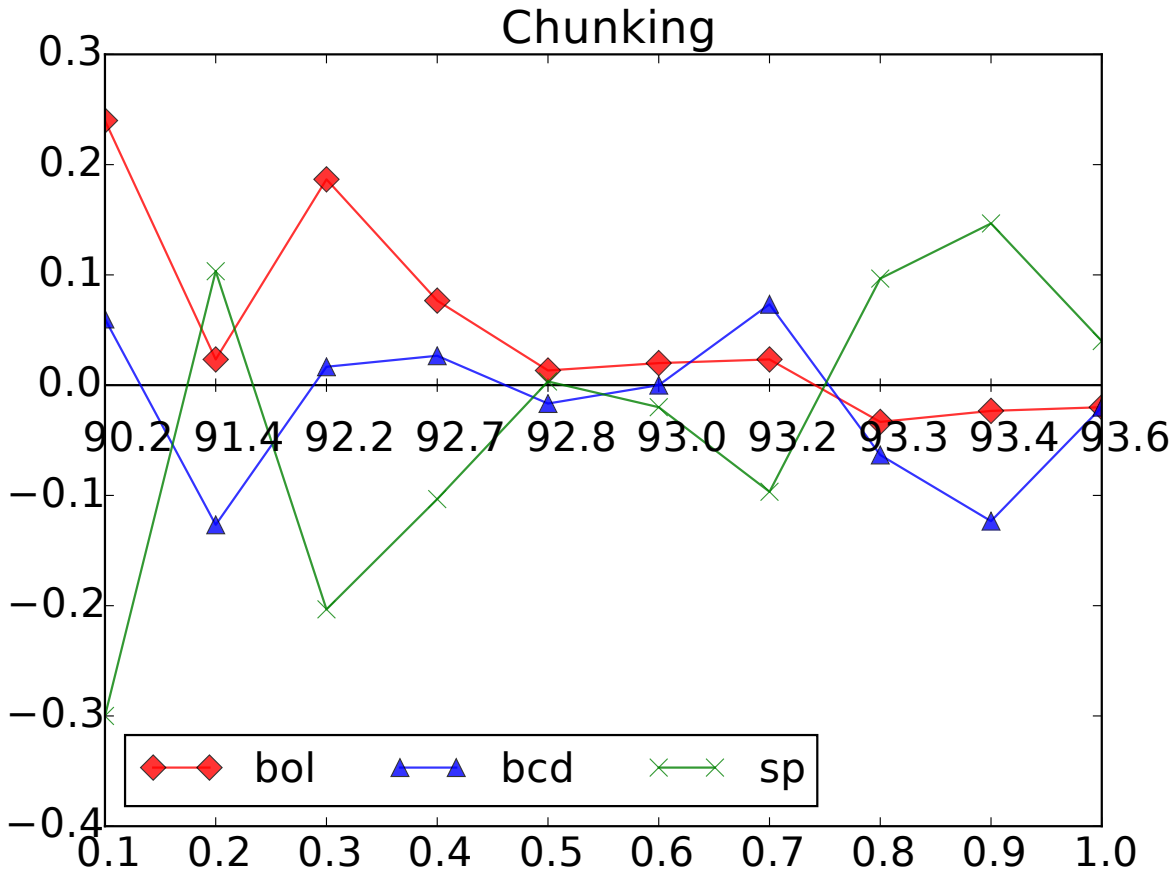


Figure: Text chunking.

Experiments

- Compared with linear models
  - When the training set is small, the advantage of bol is more obvious
  - The model is more compact
- Compared with blockwise coordinate descent
  - Prevent attracting by solutions of 0-order model.

Experiments

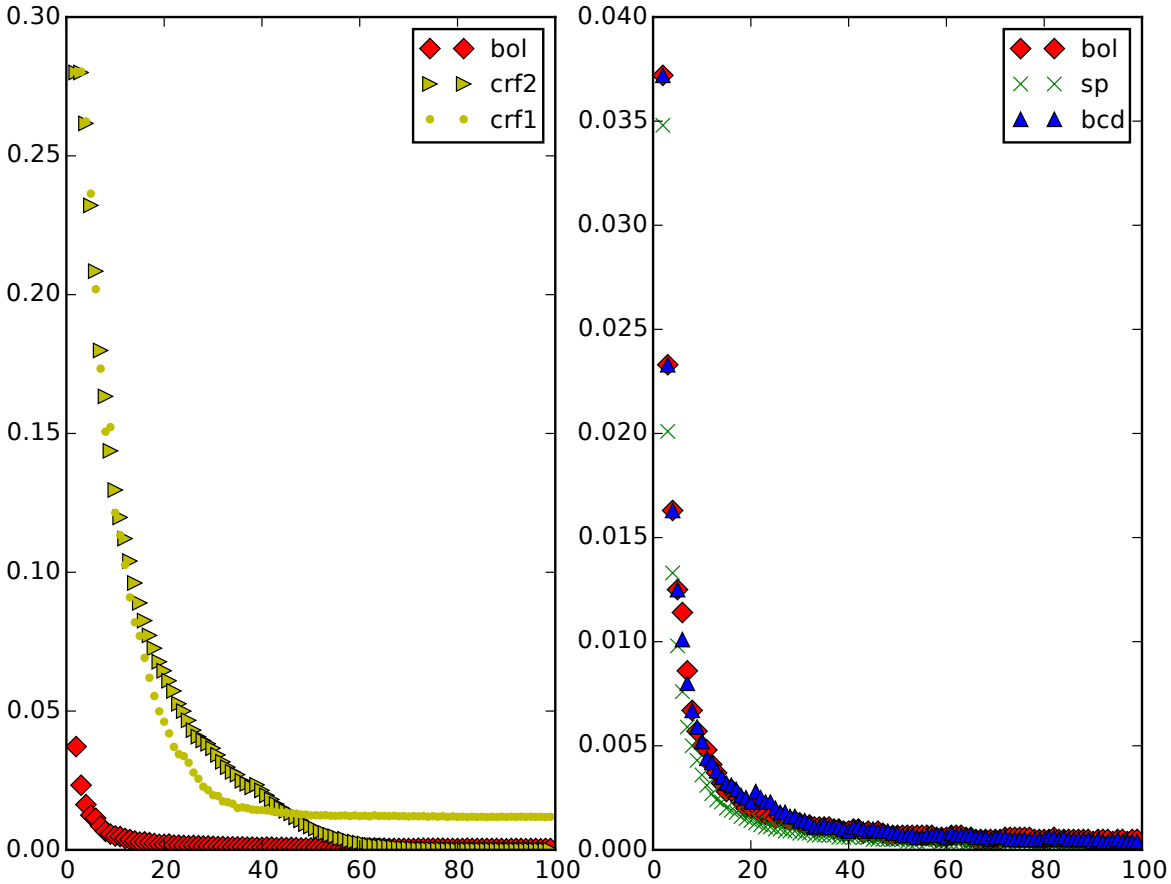


Figure: Convergence.

Conclusion

- An online learning algorithm for bilinear model
- A second order approximation of the squared spectral norm
- Future works
  - rank k constraints
  - roughly, needs to compute the leading k singular vectors

Thanks