Lecture 15 - scribbles - cutting plane

Tuesday, March 13, 2018 14:32

today: • more SVMstruct properties
• cutting plane alg.
• FW

more properties of SVMstruct dual:

$$w(\alpha) = \frac{1}{\lambda n} \sum_i \sum_{\tilde{y} \in Y_i} \alpha_i(\tilde{y}) \psi_i(\tilde{y})$$

$$gap(\alpha) = \frac{1}{n} \sum_i \left[H_i(w) - \sum_{\tilde{y} \in Y_i} \alpha_i(\tilde{y}) H_i(\tilde{y}; w) \right]$$

$w = w(\alpha)$

$\hookrightarrow \ell_i(\tilde{y}) - w^T \psi_i(\tilde{y})$

$$w^* = \frac{1}{\lambda} \left(\frac{1}{n} \sum_i \sum_{\tilde{y} \in Y_i} \alpha_i^*(\tilde{y}) \psi_i(\tilde{y}) \right)$$

Let $R_i \triangleq \max_{\tilde{y}} \|\psi_i(\tilde{y})\|_2$

$\bar{R} \triangleq \frac{1}{n} \sum_i R_i$

then 1) $\|w^*\| \le \frac{1}{\lambda} \frac{1}{n} \sum_i \left(\sum_{\tilde{y}} \alpha_i^*(\tilde{y}) \underbrace{\|\psi_i(\tilde{y})\|}_{\le R_i} \right)$

$\sum_{\tilde{y}} \alpha_i^*(\tilde{y}) = 1$

$$\le \frac{1}{\lambda} \left(\frac{1}{n} \sum_i R_i \right) = \frac{\bar{R}}{\lambda}$$

$\varphi(x^{(i)}, y^{(i)}) - \varphi(x^{(i)}, \tilde{y})$

Kernel trick: $\langle w, \varphi(x,y) \rangle = \frac{1}{\lambda n} \sum_i \sum_{\tilde{y}} \alpha_i(\tilde{y}) \langle \psi_i(\tilde{y}), \varphi(x,y) \rangle$

$k(x^{(i)}, y^{(i)}; x, y) - k(x^{(i)}, \tilde{y}; x, y)$

2) suppose rescale features $\boxed{\tilde{\psi} = b\psi}$

$\|A\alpha\|^2 \rightsquigarrow \alpha^T K \alpha$

$$\tilde{H}_c(y; w) = \ell_i(y) - \langle \tilde{w}, \tilde{\psi}_i(y) \rangle$$

$$\tilde{w} = \frac{1}{\tilde{\lambda}} \frac{1}{n} \sum_i \sum_{\tilde{y}} \{ \alpha_i^*(\tilde{y}) \tilde{\psi}_i(\tilde{y}) \} b \psi_i(\tilde{y})$$

set $\boxed{\tilde{\lambda} = b^2 \lambda}$

w^*, α^* optimal for (P)
$\Leftrightarrow$
$\tilde{w} = \frac{w^*}{b}$ and α^* optimal for $(\tilde{P})$ where $\tilde{\psi} = b\psi$

$\tilde{\lambda}$ n ; $\tilde{y}$

$\Rightarrow \tilde{w}(\alpha^*) = \frac{w^*}{b}$ and thus $\hat{w}$ is optimal for new ~~problem~~ rescaled problem

$w = \frac{w^*}{b}$ and α^* optimal for $(\hat{P})$ where $\tilde{\psi} = b\psi$, $\hat{\lambda} = b^2\lambda$

3) scale loss: $\hat{\ell} = b \cdot \ell \Rightarrow \tilde{\lambda} = \frac{\lambda}{b}$

4) etc...

M³-net example (dual):

$w(\alpha) = A\alpha = \sum_i A_i \alpha_i$

suppose $\psi(y) = \sum_c \psi_c(y_c)$

$\lambda n A_i \alpha_i = \sum_y \alpha_i(y)\, \psi_i(y) = \sum_y \alpha_i(y) \sum_c \psi_{i,c}(y_c)$

$= \sum_c \sum_{y_c} \psi_{i,c}(y_c) \left[\sum_{y' \text{ s.t. } y'_c = y_c} \alpha_i(y') \right]$ $\triangleq \mu_{i,c}(y_c)$ ← marginal variable

$\Downarrow$

$A_i \alpha_i = \tilde{A}_i \mu_i$ where $(\tilde{A}_i)_{:,(c,y_c)} = \frac{\psi_{i,c}(y_c)}{\lambda n}$

↳ #columns $\sum_c |\mathcal{Y}_c|$

similarly, suppose $\ell_i(y) = \sum_c \ell_{i,c}(y_c)$

define $\tilde{b}_{i,c}(y_c) \triangleq \frac{\ell_{i,c}(y_c)}{n}$

$\langle b_i, \alpha_i \rangle = \langle \tilde{b}_i, \mu_i(\alpha) \rangle$

$\alpha_i \in \Delta_{|\mathcal{Y}_i|} \Rightarrow \mu_i \in \mathcal{M}_i$ ← marginal polytope

tractable size QP?

thus we can replace $\max_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} -\lambda \frac{\|A\alpha\|^2}{2} + b^T\alpha$ with $\max_{\mu_i \in \mathcal{M}_i} -\lambda \frac{\|\tilde{A}\mu\|^2}{2} + \tilde{b}^T\mu$

suppose $\mathcal{M}_i$ is tractable

M³-net paper used "structured SMO" algorithm

block coordinate ascent using pair of variables

structured SVM algorithm

block-coordinate ascent using pair of variables at a time
[similar to pairwise FW]

(P) $\prod_i |\mathcal{Y}_i|$ constraints

1-slack version

$$\min_{w,\xi} \frac{\lambda}{2}\|w\|^2 + \xi \quad \text{s.t.} \quad \xi \geq \frac{1}{n}\sum_i H_i(y_i;w) \;\; \forall \tilde{y}_i \in \mathcal{Y}_i, i=1,\dots,n$$

(D)

$$\max -\frac{1}{2\lambda}\|Aa\|^2 + \langle b, a\rangle \quad a \in \Delta|\mathcal{Y}_1| \times \dots \times |\mathcal{Y}_n|$$

$w(a) = \sum \alpha(\tilde{y}_1,\dots,\tilde{y}_n)$ in $\tilde{y} \in \mathcal{Y}_1 \times \dots \times \mathcal{Y}_n$

$\sum_i |\mathcal{Y}_i|$ constraints

[ML 2009] 1-slack SVMstruct paper

$O(\frac{1}{\varepsilon})$ # iterations

big memory saving as 1 constraint per iteration

constraint generation algorithm:

[Tsochantaridis & al. JMLR 2005]

want to solve

$$\min_{w,\xi} \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_i \xi_i$$

$$\text{s.t.} \quad \xi_i \geq H_i(y_i;w) \;\; \forall y \in \mathcal{Y}_i \quad \xi_i \geq 0$$

exponential # of constraints

n-slack version

(P)

(D)

$$\max -\frac{1}{2\lambda}\|Aa\|^2 + \langle b, a\rangle \quad \text{s.t.} \quad a_i \in \Delta(|\mathcal{Y}_i|)$$

→ # variables

for LP, this is called "column generation"

SVMstruct algorithm:

"constraint generation alg." aka here as "cutting plane"

Iterate solving QPs with more and more constraints

1) start with no constraint $\Rightarrow w^{(0)} = 0$, $\xi^{(0)} = 0$

2) repeat: for each i, find $\hat{y}_i = \operatorname{argmax}_{\tilde{y} \in \mathcal{Y}_i} H_i(\tilde{y};w^{(t)})$ ← loss augmented decoding

- add $\xi_i \geq H_i(\hat{y}_i;w)$ constraint to QP } $O(n)$ step
- resolve QP(w,ξ) with these constraints to get $w^{(t+1)}, \xi^{(t+1)}$ [e.g. using CVXopt]

stop when primal-dual gap $\leq \varepsilon$

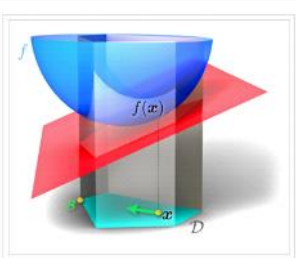
[in 2005, showed that algorithm stops after $O(\frac{1}{\varepsilon^2})$ iterations
refined later to $O(\frac{1}{\varepsilon})$]

15h40

Frank-Wolfe algorithm

for smooth constrained optimization

(motivation dual of SVM struct:
$\min_{a \in \Delta_{|S_i|}} \frac{\lambda}{2}\|Aa\|^2 - b^T a$)



A step of the Frank–Wolfe algorithm

1940s: simplex algorithm to solve LPs

1956: Marguerite Frank & Phil Wolfe

→ non-linear optimization by iterating LPs

setup: $\min_x f(x)$ s.t. $x \in M$

- f is L-smooth ie. ∇f is L-Lipschitz and f is convex
- M is convex and bounded set

(implicit assumption is that LMO is efficient)

(more generally, can get rates for f continuously differentiable)
e.g. ∇f is Hölder-cts

FW algorithm:

"linear minimization oracle"

start with $x_0 \in M$

for $t = 0, \ldots$

FW corner

compute $s_t = \underset{s \in M}{\operatorname{argmin}} \langle s, \nabla f(x_t) \rangle$ ← LMO

$f(s) \geq f(x_t) + \langle \nabla f(x_t), s - x_t \rangle \quad \forall s \in M$

minimize → linear approx of f at x_t

stopping criterion

SEM

stopping criterion

[Let $g_t \triangleq \langle s_t - x_t, -\nabla f(x_t)\rangle$ (FW gap) if $g_t \le \epsilon$; output x_t]

$x_{t+1} = (1-\gamma_t)x_t + \gamma_t s_t$ ← step-size (convex combo between x_t & s_t)

$= x_t + \gamma_t \underbrace{(s_t - x_t)}_{d_t}$ $\quad \gamma_t \in [0,1]$

end

output x_{t+1}

step-size choice: $\gamma_t =$

- universal choice $\frac{2}{t+2}$
- line-search $\gamma_t = \underset{\gamma\in[0,1]}{\operatorname{argmin}} f(x_t + \gamma(s_t - x_t))$
- adaptive: $\frac{g_t}{L\|s_t - x_t\|^2}$ or $\frac{g_t}{c_f}$ truncated at 1 (c_f affine invariant constant)

(*) big motivation for FW is that LMO is often much cheaper than projections and cheap for many sets M appearing in ML

properties:

1) $f(x_t) - \min_{x\in M} f(x) \le O(\frac{1}{t})$ ($\min_{x\in M} f(x) \triangleq f^*$)

2) FW-gap $g_t \ge f(x_t) - f^*$ → certificate of suboptimality

$\min_{s\le t} g_s \le O(\frac{1}{t})$

3) $x_t = p_0^t x_0 + \sum_{u=1}^{t} p_u^t s_{u-1}$ → x_t has "sparse" expansion in terms of the FW-corners $\{s_u\}_{u=1}^{t-1}$

where $\sum_{u=0}^{t} p_u^t = 1$ $\quad p_u^t \ge 0$

sparse method → popular in ML (*)(*)

(see here can run FW on SVM's struct dual assuming can compute LMO)

4) FW is affine covariant (like Newton)

Thm 2) gap bound: $f(s) \geq f(x_t) + \langle \nabla f(x_t), s - x_t \rangle$ $\forall s$ by convexity

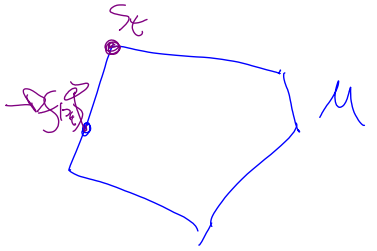
min with respect to $s \in \mathcal{M}$ on both sides

$$f^* \geq f(x_t) + \underbrace{\langle \nabla f(x_t), s_t - x_t \rangle}_{-g_t}$$

$$\Rightarrow \boxed{g_t \geq f(x_t) - f^* \geq 0}$$

interpretations: s_t is min. linear approx. of f

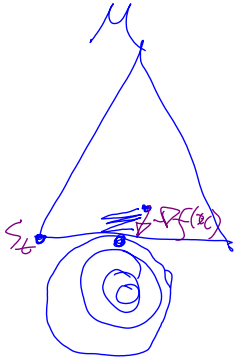
or max inner product with $-\nabla f(x_t)$



comment: FW step moves mass uniformly away from active set to FW-corner s_t

$$x_t = \sum_u \alpha_u s_u \qquad x_{t+1} = (1-\gamma_t) x_t + \gamma_t s_t$$

$$x_{t+1} = \sum_u \underbrace{\alpha_u (1-\gamma_t)}_{\text{shrinking previous coordinate}} s_u + \gamma_t s_t$$



unless step-size $\gamma_t = 1$, FW is never removing a corner from its expansion

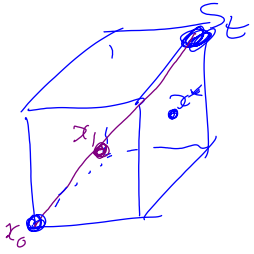
$\Rightarrow$ zig-zag close to boundary on polytopes (this is why $O(\frac{1}{t})$ rate

→ zig-zag close to boundary on polytopes

(this is why $O(\frac{1}{t})$ rate even if f is μ-strongly convex)

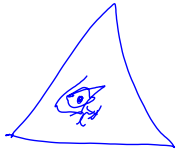
but FW has no problem — on "strongly convex sets"
↳ sublevel set of a strongly convex fct.

— when solution is in the interior



$(O(\exp(-\gamma t))$ convergence rate

$\|$)



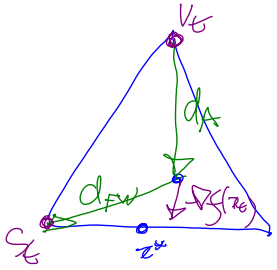
Away-step FW fix: (solve the zig-zagging problem)

in addition to compute $S_t = \underset{S \in M}{\operatorname{argmin}} \langle \nabla f(x_t), S \rangle$

$x_t = \sum_u \alpha_u S_u$

active-set$(x_t) = \{S_u : \alpha_u > 0\}$

compute $V_t = \underset{S \in \text{active-set}(x_t)}{\operatorname{argmax}} \langle \nabla f(x_t), S \rangle$



$d_{FW} \triangleq S_t - x_t$

$d_A \triangleq x_t - V_t$

AFW picks the direction with best inner-product

i.e. pick d_A if $\langle d_A, \nabla f(x_t) \rangle > \langle d_{FW}, \nabla f(x_t) \rangle$

If use d_A, let $\gamma_t = \underset{\gamma}{\operatorname{argmin}} f(x_t + \gamma d_A)$

$\gamma \in [0, \gamma_{max}]$ γ_{max} is determined from α_u's

$$x_{t+1} = x_t + \gamma_t (x_t - v_t)$$

$$x_{t+1} = \sum_u (1+\gamma_t)\alpha_u s_u - \gamma_t v_t$$

$$\gamma - \gamma\alpha = \alpha$$

let α be coefficient of v_t; $(1+\gamma_{max})\alpha - \gamma_{max} = 0$

$$\boxed{\gamma_{max} = \frac{\alpha}{1-\alpha}}$$

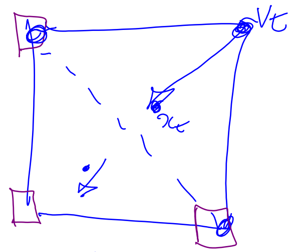
when $\gamma_t = \gamma_{max}$

"drop step" → removed v_t from expansion

run AFW: either you maintain some expansion of $x_t = \sum_u \mu_u \tilde{s}_u$

or you have a feasibility oracle + away-step oracle
[see NIPS 2016 paper]
Meshi & Garber → here
$\text{corners}(M) \subseteq \{0,1\}^d$

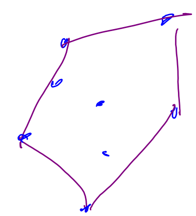
$M \rightarrow x \geq 0$ and $Ax = b$



AFW has linear convergence rate on polytopes when f is strongly convex

(*) other comments: if $M = \text{conv}(\mathcal{A})$ where $\mathcal{A}$ is some finite set "atoms"

$$\text{LMO}(r): \min_{s \in M = \text{conv}(\mathcal{A})} \langle s, r \rangle = \min_{a \in \mathcal{A}} \langle a, r \rangle$$



→ lots of application of this in ML ...

$S \in M = \text{conv}(A)$ $a \in A$

↳ lots of application of this in ML

e.g. • $A \to$ integer flows $\quad$ LMO → min cost network flow
$\text{conv}(A) \to$ flow polytope

• $A \to$ digit assignment in graph $\quad$ LMO → max-product or graph cut alg. for submodular potential (see later)