Lecture 12 - structured SVMs

Thursday, February 22, 2024 9:33 AM

today: structured SVM optimization

other approaches to optimize SVMstruct

(UP) unconstrained primal

$$\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^n H_i(w)$$

[convex, unconstrained, non-smooth]

(PQP) primal QP → quadratic problem

$$\min_{w,(\xi_i)_{i=1}^n} \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \xi_i$$

[constrained formulation]

$$\xi_i \geq H_i(w,\tilde{y}) \quad \forall \tilde{y}\in\mathcal{Y}_i \;\&\; \forall i$$

(smooth) convex QP with exp. # of linear constraints

1) generic approach to use convexity of loss augmented decoding: [Taskar & al. ICML 2005]

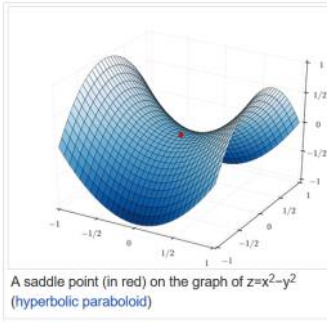
idea: here, we suppose that loss-augmented decoding can be expressed as a "compact" maximization problem of a concave fct.

i.e. $H_i(w) = \max_{\tilde{y}\in\mathcal{Y}_i} \ell_i(\tilde{y}) - \langle w, \psi_i(\tilde{y})\rangle = \max_{z\in\mathcal{Z}} g_i(w,z)$

$\mathcal{Y}_i$: discrete; z: cts. variable; $\mathcal{Z}$: convex set

where g_i is concave in z and convex in w

$\mathcal{Z}$: • should not depend on w • ideally have a tractable description



A saddle point (in red) on the graph of $z=x^2-y^2$ (hyperbolic paraboloid)

a) saddle point formulation:

$$\min_w \max_{\substack{z_i\in\mathcal{Z}_i \\ i=1,\dots,n}} \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^n g_i(w,z_i)$$

$$\min_w \underbrace{\max_z f(w,z)}_{m(w)}$$

f convex in w, concave in z

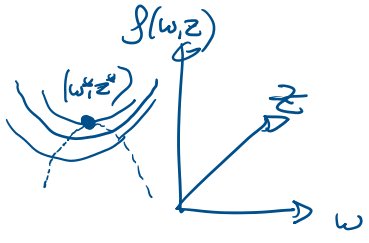
convex-concave saddle pt. problem

(under reg. conditions $\min\max = \max\min$ → "saddle point")

$$\forall z\; f(w^*,z) \leq f(w^*,z^*) \leq f(w,z^*)\;\forall w$$

$$w^* \in \operatorname*{argmin}_w f(w,z^*)$$

$$z^* \in \operatorname*{argmax}_z f(w^*,z)$$



in general:

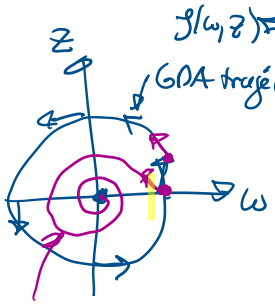
$$\min_w \max_z f(w,z) \geq \max_z \min_w f(w,z)$$

$z^* \in \arg\max_z f(w^*, z)$

→ might not exist ?

[but it always exists for convex-concave + reg. condition]
e.g. bounded convex constraints

$f(w, z) = \langle w, z \rangle$



$\begin{pmatrix} \nabla_w f \\ \nabla_z f \end{pmatrix} = \begin{pmatrix} z \\ -w \end{pmatrix}$

Standard alg. :

extragradient alg. → converges $(\frac{1}{t})$ for convex-concave games

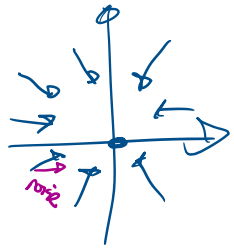
$$\begin{pmatrix} \tilde{w}_{t+1} \\ \tilde{z}_{t+1} \end{pmatrix} = \begin{pmatrix} w_t \\ z_t \end{pmatrix} + \gamma_t \begin{pmatrix} -\nabla_w f(w_t, z_t) \\ \nabla_z f(w_t, z_t) \end{pmatrix}$$

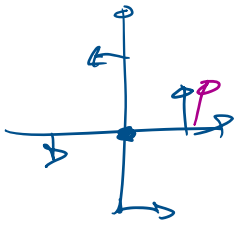
"lookahead step"

$$\begin{pmatrix} w_{t+1} \\ z_{t+1} \end{pmatrix} = \begin{pmatrix} w_t \\ z_t \end{pmatrix} + \gamma_t \begin{pmatrix} -\nabla_w f(\tilde{w}_{t+1}, \tilde{z}_{t+1}) \\ \nabla_z f(\text{"} \text{"}) \end{pmatrix}$$

extragradient alg.

stoch. min vs stochastic games





$\tilde{x}_{t+1} = x_t - \gamma_t F(x_t)$ $\quad x = \begin{pmatrix} w \\ z \end{pmatrix}$

$x_{t+1} = x_t - \gamma_t F(\tilde{x}_{t+1})$ $\quad F(x) = \begin{pmatrix} \nabla_w f \\ -\nabla_z f \end{pmatrix}$

1st order approx. to the implicit method

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

$x^* = x^* - \gamma F(x^*)$

extragradient to structured SVM [Taskar & al. JMLR 2006]

Ch33

b) small "complicated" QP formulation (for structured SVM)

convex dual of $\max_{z_i \in Z_i} g_i(w; z_i)$

$$H_i(w) = \max_{z_i \in Z_i} g_i(w; z_i) = \min_{v_i \in V_i(w)} \tilde{g}_i(w; v_i)$$

using strong duality — dual variables

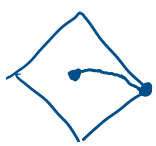
obtain $\min_w \min_{v_i \in V_i(w) \ \forall i} \frac{\lambda \|w\|^2}{2} + \frac{1}{n} \sum_{i=1}^n \tilde{g}_i(w; v_i)$

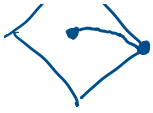
if $\tilde{g}_i$ is jointly convex in w & v_i we get a "tractable" convex min.

assume $v_i \in V_i(w)$ ⇒ convex constraint in w

→ can solve with favorite convex min. alg.

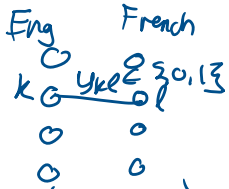
if d & n not too big, use interior point solver
e.g. Mosek, Cplex (commercial)
CVXopt (free, python)





examples $g_i(w_i, z_i)$

I) word alignment:



features on pairs (x_k^E, x_ℓ^F)

recall that score $s(x, y; w) = \sum_{k,\ell} y_{k\ell} \left[w^T \varphi(x_k^E, x_\ell^F) \right]$

Let $y \in \{0,1\}^{L^E \cdot L^F}$

Let matrix F be $d \times L^E \cdot L^F$ $\quad \left| \left[\cdots \varphi_{k\ell} \cdots \right] \right.$

$s(x, y; w) = w^T F y$

$s(x^{(i)}, \tilde{y}; w) = w^T F_i \tilde{y}$

decoding: $h_w(x^{(i)}) = \underset{\tilde{y} \in \mathcal{Y}_i}{\operatorname{argmax}} \; s(x^{(i)}, \tilde{y}; w)$ becomes $\max_{\substack{y_{k\ell} \in \{0,1\} \\ y \in M_i}} w^T F_i y$ linear integer program

matching constraints

$M_i \triangleq \{ y \in \mathbb{R}^{L_E^{(i)} \cdot L_F^{(i)}} : \begin{cases} \sum_\ell y_{k\ell} = 1 \\ \sum_k y_{k\ell} \le 1 \\ 0 \le y_{k\ell} \le 1 \end{cases} \}$

# of constraints (i) $L_E + L_F^{(i)}$

$y_{11}\ y_{12}\ y_{21}\ y_{22}$

$A: \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ & & I & \end{pmatrix}$

$M_i = \{ z : Az \le \mathbb{1}, z \ge 0 \}$

$(2L + L^2) \times L^2$ matrix

(!) here, turns out that can remove integer constraint to get a "relaxed LP" give the same opt. obj. value i.e. relaxation is tight

[actually here, $M_i = \text{convex-hull}(\mathcal{Y}_i)$]

why is relaxation tight?

write $z \in M_i$ as $Az \le b_i$, $z \ge 0$; matrix A here is "totally unimodular"

which means any subdeterminant of A has value $\{+1, -1, 0\}$

$\Rightarrow$ that if b has integer entries, then all vertices of $\{z : Az \le b, z \ge 0\}$ have integer coordinates

$\Rightarrow$ relaxation is tight for any linear cost

$\Rightarrow$ relaxation is tight for any linear cost

idea: $\tilde{A}z \le \tilde{b}$, a corner of this polytope is obtained by solving

$\tilde{A}_{I:}\tilde{z} = \tilde{b}_I$ for $\tilde{A}_{I:}$ is invertible

$|I| = \dim(\tilde{z})$ $\Downarrow$

$\tilde{z}^* = \tilde{A}_{I:}^{-1} b_I$

use Cramer's rule ratio of subdeterminant

$\Rightarrow$ get integer values //

Conclusion: can write decoding as $\max_{z \in M_i} w^T F_i z$

what about loss?

Hamming loss example

$$l(y, \tilde{y}) = \sum_{k,\ell} \mathbb{1}\{y_{k\ell} \neq \tilde{y}_{k\ell}\}$$

$$= \sum_{k,\ell} (y_{k\ell} - \tilde{y}_{k\ell})^2$$

$$(y_{k\ell}^2 - 2\tilde{y}_{k\ell} y_{k\ell} + \underbrace{\tilde{y}_{k\ell}^2}_{=\tilde{y}_{k\ell}}) = (y_{k\ell}^2 + (1 - 2y_{k\ell})\tilde{y}_{k\ell})$$

cool trick:

$y^2 = y$ when $y \in \{0,1\}$

$$l_i(\tilde{y}) = \underbrace{a_i}_{\sum_{k,\ell} y_{k\ell}^{(i)}} + \underbrace{(1 - 2y^{(i)})}_{c_i}{}^T \tilde{y}$$

loss augmented decoding

$$\max_{\substack{\tilde{y}_{k\ell} \in \{0,1\} \\ \tilde{y} \in M_i}} \underbrace{a_i + c_i^T \tilde{y}}_{l_i(\tilde{y})} - (\underbrace{w^T F_i y^{(i)}}_{w^T \psi_i(y)} - w^T F_i \tilde{y})$$

$$a_i - w^T F_i y^{(i)} + \max_{\substack{\tilde{y}_{k\ell} \in \{0,1\} \\ \tilde{y} \in M_i}} (F_i^T w + c_i)^T \tilde{y}$$

$$= \max_{z \in M_i} \underbrace{(F_i^T w + c_i)^T z + a_i - w^T F_i y^{(i)}}_{g_i(w,z)}$$

LP duality:

$$\max_{\substack{Az \le b_i \\ z \ge 0}} \tilde{c}_i^T z = \min_{\substack{A^T v \ge \tilde{c}_i \\ v \ge 0}} b_i^T v$$

$$\max_{z \in Z_i} g_i(w,z) = \min_{v \in V_i(w)} \hat{g}_i(v; z)$$

here $\hat{c}_i \triangleq F_i^T w + c_i$

$z \geq 0$ $\quad v \geq 0$

here $\hat{c}_i \triangleq F_i^T w + c_i$

A_i is $2L \times L^2$

(note $b_i = -1$ here)

Sum struct obj. becomes (small QP)

$$\min_w \min_{\{v_i\}_{i=1}^n} \frac{1}{2}\lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^n [a_i - w^T F_i y^{(i)} + b_i^T v_i]$$

$$\text{s.t. } A_i^T v_i \geq F_i^T w + c_i \qquad \text{"small complicated QP"}$$
$$v_i \geq 0$$

compare with saddle pt. formulation

$$\min_w \max_{\{z_i\}_{i=1}^n} \frac{1}{2}\lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^n [a_i - w^T F_i y^{(i)}] + [(F_i^T w + c_i)^T z_i]$$

$z_i \in M_i$ → simpler constraints