Lecture 12 - structured SVMs

Thursday, February 23, 2023 1:29 PM

today: structured SVM optimization

other approaches to optimize SVM struct

(UP) unconstrained primal

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

[unconstrained] non-smooth

(PQP) primal QP — quadratic program

$$\boxed{\begin{aligned}&\min_{w,(\xi_i)_{i=1}^n} \frac{\lambda\|w\|^2}{2} + \frac{1}{n}\sum_i \xi_i \\ &\xi_i \geq H_i(w;\tilde{y}) \quad \forall \tilde{y}\in \mathcal{Y}_i,\ \forall i\end{aligned}}$$

[constrained formulation

(smooth) strongly 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" concave maximization problem of a concave obj. over convex constraints

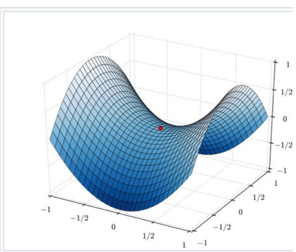
i.e. $H_i(w) = \max_{\tilde{y}\in\mathcal{Y}_i} \ell_i(\tilde{y}) - \langle w, \psi_i(\tilde{y})\rangle = \max_{z\in Z} g_i(w,z)$

$\tilde{y}\in\mathcal{Y}_i$: discrete

$z \in Z$ cts. convex set

where g_i is concave in z and convex in w

Z • should not depend on w
• and have a tractable description



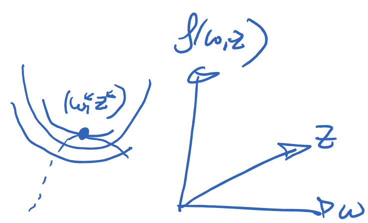
A saddle point (in red) on the graph of $z=x^2-y^2$ (hyperbolic paraboloid)

a) saddle pt. formulation:

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

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

• convex in w
• concave in z



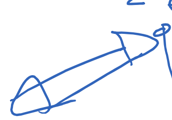
convex-concave saddle pt. problem

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

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

(under reg. conditions min-max = max-min) → "saddle pt."

"zero-sum" game



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

might not exist;

$\forall z \quad f(w^*, z) \le f(w^*, z^*) \le f(w, z^*) \quad \forall w$

might not exist;

[but always exists for convex-concave obj. + reg. condition]
↳ e.g. bounded convex constraint sets

in general though:

min max $\ge$ max min

$$\min_w \max_z f(w,z) \ge \max_z \min_{w'} f(w', z)$$

$$\min \ (x-a)^2 \quad \text{s.t.} \quad x^2 \le 0 \qquad f(x,\lambda) = (x-a)^2 + \lambda x^2$$

$$\min_x \rightarrow \quad 2(x-a) + \lambda 2x = 0$$

$$(2+2\lambda)x = 2a$$

$$\Rightarrow \quad x^* = \frac{a}{1+\lambda}$$

$$d(\lambda) = f(x^*(\lambda), \lambda) = a^2\left(\frac{\cancel{1+\lambda} - 1}{1+\lambda}\right)^2 + \lambda \frac{a^2}{(1+\lambda)^2}$$

$$= \frac{a^2}{(1+\lambda)^2}\left[\lambda^2 + \lambda\right] = \frac{a^2 \lambda}{1+\lambda}$$

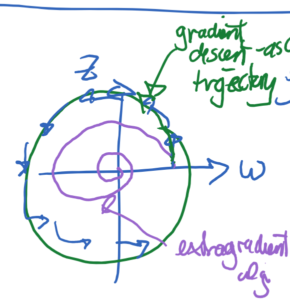
$$\frac{d}{d\lambda}\left(\frac{\lambda}{1+\lambda}\right) = \left[\frac{1(1+\cancel{\lambda}) - \cancel{\lambda}}{(1+\lambda)^2}\right] = \frac{1}{(1+\lambda)^2}$$

$$\lambda^* = +\infty$$

$$d(\lambda^*) = \max_{\lambda \ge 0} \min_x f(x,\lambda) = 0 \qquad \min_x \max_{\lambda \ge 0} f(x,\lambda) = a^2$$

Standard alg.:

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



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

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

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

"lookahead step"

$$\begin{pmatrix} w_{t+1} \\ z_{t+1} \end{pmatrix} = \begin{pmatrix} w_t \\ z_t \end{pmatrix} + \gamma_t \begin{pmatrix} -\nabla_w f(\tilde{w}_{t+1}, \tilde{z}_{t+1}) \\ \nabla_z f('' , '') \end{pmatrix}$$

$$x = \begin{pmatrix} w \\ z \end{pmatrix}$$

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

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

1st order approximation to the implicit method

$$x_{t+1} = x_t - \gamma_t F(x_{t+1}) \qquad x^* = x^* - \gamma F(x^*)$$

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

14h35 $\quad [\ f(w,z) = f(w) - g(z)\ ]$

b) small "complicated" QP formulation (for SVM struct) → convex dual of $\max_z g_i(w,z)$

$$H_i(w) = \max_{z_i \in Z_i} g_i(w, z_i) \underset{\text{strong duality}}{=} \min_{v_i \in V_i(w)} \tilde{g}_i(w; \underbrace{v_i}_{\text{dual variables}})$$

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

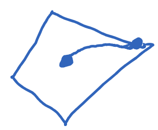
if $\tilde{g}_i$ is jointly convex in w & v_i

we get a "tractable" convex min. problem

→ can solve with favorite convex min. alg.

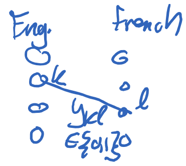
if d & n is not too big, use interior point method solver

e.g. Mosek, Cplex (commercial)
CVX opt. (free, python)



examples $g_i(w; z_i)$

I) word alignment :



features on pair (x_k^E, x_l^F)

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

let $y \in \{0,1\}^{L_E \times L_F}$

let matrix F be $d \times L_E \cdot L_F$ $\quad \left[\;—\; \varphi_{kl} \;\; \right]$

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

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

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

$$M_i \triangleq \left\{ y \in \mathbb{R}^{L_E^{(i)} \times L_F^{(i)}} : \left.\begin{array}{l} \sum_l y_{kl} \leq 1 \;\; \forall k \\ \sum_k y_{kl} \leq 1 \;\; \forall l \end{array}\right\} \text{matching constraints}, \;\; 0 \leq y_{kl} \leq 1 \right\}$$

$$M_i = \left\{ z : \begin{array}{l} Az \leq \mathbb{1} \\ z \geq 0 \end{array} \right\}$$

$$A : \begin{array}{cccc} y_{11} & y_{12} & y_{21} & y_{22} \end{array}$$
$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

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

$z \geq 0$

constraints $L_E^{(i)} + L_P^{(i)}$ $\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ & I & & \end{pmatrix}$

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

⊛ here, turns out that can remove the integer constraint to get a "relaxed LP" give same opt. obj. value

i.e. relaxation is tight

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

why is relaxation tight?

write $z \in M_i$ as $\begin{array}{l} Az \leq b_i \\ z \geq 0 \end{array}$; matrix A here is "totally unimodular"

which means that subdeterminant of A has value $\begin{cases} +1 \\ 0 \\ -1 \end{cases}$

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

$\Rightarrow$ relaxation is tight for any linear cost

idea: $\tilde{A} z \leq \tilde{b}$, a corner of this is obtain by solving

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

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

$\Downarrow$

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

Cramer's rule

$\rightarrow$ ratio of subdeterminant

$\Rightarrow$ get integer values //

conclusion: can write decoding as $\max\limits_{z \in M_i} w^T F_i z$

what about loss?

Hamming loss example

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

$$\ell(y, \tilde{y}) = \sum_{k,l} \underbrace{\mathbb{1}\{y_{kl} \neq \tilde{y}_{kl}\}}_{(y_{kl} - \tilde{y}_{k,l})^2}$$

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

$\tilde{y}_{kl}^2 = \tilde{y}_{kl}$

$\sum_{k,l} y_{kl}^2$

$$\ell_i(\tilde{y}) = a_i + \underbrace{(1 - 2y^{(i)})}_{c_i}{}^T \tilde{y}$$

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

$$\ell_i(y) = a_i + \underbrace{[1 - 2y^{(i)}]^T}_{c_i} y$$

<u>loss augmented decoding</u>

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

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

<u>LP duality:</u>

$$\max_{\substack{A_i z \leq b_i \\ z \geq 0}} \tilde{c}_i^T z = \min_{\substack{A_i^T v \geq \tilde{c}_i \\ v \geq 0}} b_i^T v \qquad\qquad \max_{z \in M_i} g_i(w; z) = \min_{v \in V_i(w)} \hat{g}_i(w; z)$$

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

A_i is $(2L + L^2) \times L^2$

SVM struct obj. becomes (small QP)

$$\boxed{\begin{aligned} &\min_{w} \min_{\{v_i\}_{i=1}^n} \frac{1}{2} \lambda \|w\|^2 + \frac{1}{n} \sum_{i=1}^n \left[a_i - w^T F_i y^{(i)} + b_i^T v_i\right] \\ &\text{s.t. } A_i^T v_i \geq F_i^T w + c_i \\ &\qquad v_i \geq 0 \end{aligned}}$$

"small complicated QP"

Compare with saddle pt. formulation

$$\boxed{\min_{w} \max_{\substack{\{z_i\}_{i=1}^n \\ z_i \in M_i}} \frac{1}{2} \lambda \|w\|^2 + \frac{1}{n} \sum_{i=1}^n \left[a_i - w^T F_i y^{(i)}\right] + \left[(F_i^T w + c_i)^T z_i\right]}$$

simpler constraints