Lecture 12 - structured SVM

Thursday, February 13, 2020 13:37

today: structured SVM optimization

other approaches to optimize SVM struct

(UP) unconstrained primal

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

[unconstrained] non-smooth

(PQP) primal QP → quadratic program

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

[constrained formulation]

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

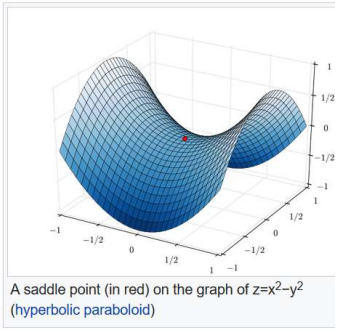
(smooth) convex QP with exp. # of linear constraints

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

ℓ_i discrete; $\mathcal{Z}_i$ convex set; g_i where g_i is concave in z and convex in w



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

$\mathcal{Z}_i$:
- should not depend on w
- we have tractable description

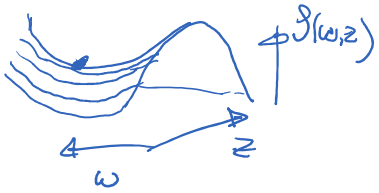
a) saddle point formulation:

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

$$\min_w \max_z \mathcal{L}(w,z)$$

- convex in w
- concave in z

convex-concave saddle point problem



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

In general: $\min_w \max_z \geq \max_z \min_w$

$$\forall z \quad \mathcal{L}(w^*,z) \leq \mathcal{L}(w^*,z^*) \leq \mathcal{L}(w,z^*) \quad \forall w$$

circular dependence: $w^* \in \operatorname{argmin}_w \mathcal{L}(w,z^*)$, $z^* \in \operatorname{argmax}_z \mathcal{L}(w^*,z)$

dependance $\rightarrow$ $z^* \in \underset{z}{\operatorname{argmax}}\, g(w^*, z)$

$\Rightarrow$ might not exist ? [but always exist for convex-concave + reg. conditions e.g. 1-coercive]

Standard alg: $\rightarrow$ converges $O(\frac{1}{t})$ for convex-concave game

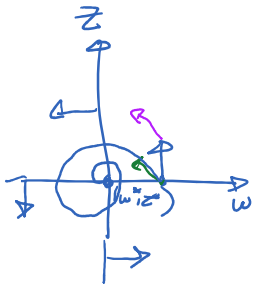
extragradient algorithm : (lookahead step)

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

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

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

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



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

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

1st-order approx. to implicit method:

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

applied to structure SVM [Taskar et al. JMLR 2006]

b) Small "complicated" Opt formulation (for structured SVM)

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

$\tilde{g}_i$: convex dual of $\max g_i(w; z_i)$ $z_i \in Z_i$

use strong duality

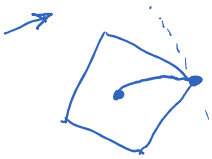
v_i: dual variables

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

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

$\rightarrow$ can solve with generic convex min alg

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



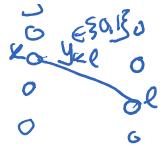
14h35

Examples of $g_i(w; z_i)$

I) word alignment :

$\rightarrow$ features on pairs (E_i, F_j)

I) word alignment :



features on pair (x^E_k, x^F_ℓ)

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

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

let matrix F be $\left[\cdots \varphi_{k\ell} \cdots \right]$ $d \times (L_E \cdot L_F)$

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

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

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

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

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

y_{11} y_{12} y_{21} y_{22}

$A: \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$ $M_i = \{z : Az \le b, z \ge 0\}$

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

ie. relaxation is tight

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

reasons that relaxation is tight

a) write $z \in M_i$ as $Az \le b_i$, $z \ge 0$ matrix A here is "totally modular"

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

idea : $\tilde{A}\tilde{z} \le \tilde{b}$, a corner of this obtained by solving $\tilde{A}_{I:} \tilde{z} = \tilde{b}_I$ for $\tilde{A}_{I:}$ invertible

$|I| = \dim(z)$ $\quad \tilde{z} = \tilde{A}_I^{-1} b_I$

↑ Cramer's rule : ratio of substitution $\Rightarrow$ integer

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

What about loss?

Hamming loss example

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

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

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

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

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

loss-augmented decoding :

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

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

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

LP duality :

$$\max_{\substack{A_i z \leq b_i \\ z \geq 0}} \tilde{c}_i^T z = \min_{\substack{A_i^T v \geq \tilde{c}_i \\ v \geq 0}} b_i^T v$$

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

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

A_i is $2L \times L^2$

SVM struct objective becomes

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

s.t. $A_i^T v_i \geq F_i^T w + c_i$ "small complicated QP"

$v_i \geq 0$

$A'\begin{pmatrix} w \\ v \end{pmatrix} \leq b'$

compare with saddle pt. formulation

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

$z_i \in M_i$ ← simpler constraints