Lecture 18 - BCFW

Thursday, March 19, 2020 13:28

today : · FW & FCFW on SVMstruct

· BCFW

Continue FW for SVM struct

FW on dual is equivalent to batch subgradient update on primal with step-size relationship

$\alpha^{(t)} \longrightarrow w^{(t)} = A\alpha^{(t)}$ $\qquad \gamma^{(FW)} = \lambda\beta^{(\text{subgradient})}$

recall : subgradient method converges of $O(\frac{1}{t})$

when step-size $\beta_t \leq \frac{1}{\mu} O(\frac{1}{t})$

FW method with fixed step-size schedule

$\gamma_t = \frac{2}{t+2}$

gives $d(\alpha^*) - d(\alpha^{(t)}) \leq \frac{2C_f}{t+2}$

* FW-gap :

$$\langle -\nabla f(\alpha^{(t)}), s^{(t)} - \alpha^{(t)} \rangle$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[H_i(\hat{y}_i^{(t)}; w^{(t)}) - \sum_{\tilde{y}\in Y_i} \alpha_i^{(t)}(\tilde{y})\, H_i(\tilde{y}; w^{(t)}) \right]$$

$$= p(w(\alpha^{(t)})) - d(\alpha^{(t)})$$ [Lagrangian gap of lecture 15]

($w(\alpha^{(t)}) = A\alpha^{(t)}$)

here, FW-gap = Lagrangian gap on $(w(\alpha^{(t)}), \alpha^{(t)})$

[note: FW gap is not always Lag. gap]
e.g. CRF objective (dual)

* recall, that

$$\min_{s\leq t} \underbrace{g_s}_{\text{FW gap}} \leq 3 \cdot \frac{2C_f}{t+2}$$ $\qquad p(w^*) = d(\alpha^*)$

$\Rightarrow$ guarantees on $\underbrace{p(w(\alpha^{(t)}))}_{\text{primal subopt.}} - p(w^*) + \underbrace{d(\alpha^*) - d(\alpha^{(t)})}_{\text{dual subopt.}}$

also, $g^{FW}(\hat{x}_{WA}^{(t)}) \leq 3 \cdot \frac{2C_f}{t+2}$

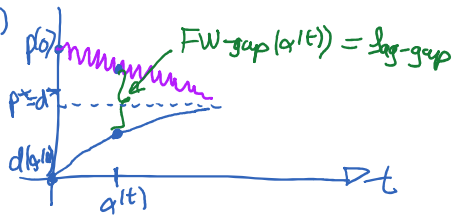
($\hat{x}_{WA}^{(t)}$: weighted avg.)

when $g^{FW}(\cdot)$ is convex [this is case when f is a quadratic]

$p(w(\alpha^{(0)})) = p(0) = \frac{1}{n}\sum_{i=1}^{n} \max_{\tilde{y}} l_i(\tilde{y})$ $\qquad p(\alpha)$ ..... FW-gap$(\alpha^{(t)})$ = lag-gap

$p(w(\alpha^{(0)})) = p(0) = \frac{1}{n}\sum_{i=1}^{n} \max_{\tilde{y}} l_i(\tilde{y}) = L_{\max}$

$d(\alpha^{(0)}) = 0$

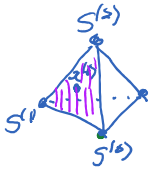


FCFC "fully corrective FW" variant

FW corners

algorithm: re-optimize f over conv-hull $(\{s^{(u)}\}_{u=0}^{t})$

→ think of it as doing "fancy line-search" on the "correction polytope"



note: could use AFW to do correction step opp.

(see "barrier FW" paper)

[special case: min norm pt. alg. MNP] → sequence of affine projections + line search

→ state of the art alg. for submodular opt.

* turns out that (batch) FCFW on dual structure is equivalent to the constraint generation / cutting plane alg. on the 1-slack formulation (primal)

why? every $s^{(t)}$ corresponds $(\hat{y}_i^{(t)})_{i=1}^{n}$

$\alpha \in \text{conv}(\{s^{(u)}\}_{u \le t})$ $\alpha = \sum_{u \le t} \tilde{\alpha}_u s^{(u)}$

g strongly convex $g(Ax) + b^T x$ "gene. strongly convex"

$w = A\alpha = \sum_{u \le t} \tilde{\alpha}_u \left(\frac{1}{\lambda n} \sum_{i=1}^{n} M_i(\hat{y}_i^{(u)}) \right)$

to summarize:

	primal problem		dual problem	complexity of alg. in # oracle calls
$O(n)$ oracle calls per iteration	batch subgradient $\beta_t = \frac{\alpha_t}{\lambda}$	⟺	batch FW (can do line search to get α_t^*)	$O\left(n \frac{R^2}{\lambda\varepsilon}\right)$
	1-slack cutting plane	⟺	FCFW	$O\left(n n^{\nu} \log\left(\frac{1}{\varepsilon}\right)\right)$ $\nu \gg 1$? (open problem) → related to pyramidal widths of marginal polytope
$O(1)$	stochastic subgradient method (no line search)	⟺ (not equivalent)	BCFW (line search)	$O\left(\frac{1}{\lambda\varepsilon} R^2\right)$

[45]

Block-coordinate optimization

"huge scale" optimization → [Nesterov 2010–2012]

proposed: randomized block-coordinate projected gradient method

setup: $\min f(x)$

s.t. $x \in \bigtimes_{i=1}^{n} M_i$

$x = (x_1, \dots, x_n)$ (block)

algorithm: pick at random i_t

then let $\begin{cases} x_{i_t}^{(t+1)} = \text{Proj}_{M_{i_t}}\left(x_{i_t}^{(t)} - \frac{1}{L_{i_t}} \nabla_{i_t} f(x^{(t)})\right) \\ x_j^{(t+1)} = x_j^{(t)} \quad \forall j \neq i \end{cases}$

(L_{i_t}: Lipschitz const. of $\nabla_{i_t} f(x^{(t)})$)

[only update block i_t at iteration t]

Nesterov showed (uniform sampling) $\mathbb{E} f(x^{(t)}) - f^* \leq \frac{2}{t+4} \left[\frac{1}{n}\sum_{i=1}^{n} L_i\right] \|x_0 - x^*\|^2$

(for convex f with L_i-Lips. gradient per block)

Block-coordinate FW (BCFW): idea: do a FW step on block i

alg.: for $t = 0, \dots$

pick i unif. at random from 1 to n

let $s_i^{(t)} = \underset{s_i \in M_i}{\text{argmin}} \langle s_i, \nabla_i f(x^{(t)}) \rangle$ [FW corner for block i]

$S_{[i]} = \begin{pmatrix} 0 \\ 0 \\ s_i \\ 0 \\ 0 \end{pmatrix}$

$\begin{cases} x_i^{(t+1)} = x_i^{(t)}(1-\gamma_t) + \gamma_t s_i^{(t)} \\ x_j^{(t+1)} = x_j^{(t)} \quad \forall j \neq i \end{cases}$

$\gamma_t =$ — line search $\underset{\gamma \in [0,1]}{\text{argmin}} f(x^{(t)} + \gamma(S_{[i]}^{(t)} - x_{[i]}^{(t)}))$

— $\frac{2n}{t+2n}$ (# of blocks)

(*) an important property: FW-gap $= \max_{s \in M} \langle -\nabla f(x), s - x \rangle \overset{\text{using product structure}}{=} \sum_i \max_{s_i \in M_i} \langle -\nabla_i f(x), s_i - x_i \rangle$

$\triangleq g_i(x)$ "block FW gap"

$$g^{FW}(x) = \sum_i g_i^{FW}(x)$$

↳ motivated "gap sampling" variant [Osokin & al. ICML 2016]

as before, can show that $g_i(x) \geq f(x) - \min_y f(y)$ s.t. $y_{[i]} = x_{[i]}$