Lecture 12 - scribbles - FW for SVMstruct

Wednesday, February 15, 2017
10:45

today: FW & BCFW for structured SVM!

minus dual objective $-d(\alpha)$

recall: batch FW on $\min_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} \underbrace{\frac{\lambda}{2}\|A\alpha\|^2 - b^T\alpha}$

$(\nabla f(\alpha))_{i,y} = -\frac{1}{n} H_i(y; w(\alpha))$ $\quad w(\alpha) = A\alpha$

$= (\lambda A^T A - b)_{i,y}$

FW step on $M = \bigoplus_{i=1}^n \Delta_{|\mathcal{Y}_i|}$

loss-augmented inference for each i: $s_i^{(t)} = e_{\hat{y}_i^{(t)}}$ where $\hat{y}_i^{(t)} = \arg\max_{\tilde{y} \in \mathcal{Y}_i} H_i(\tilde{y}; w^{(t)})$

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

from $\alpha^{(t+1)} = (1-\gamma)\alpha^{(t)} + \gamma s^{(t)}$

$w^{(t+1)} = (1-\gamma) w^{(t)} + \gamma \frac{1}{\lambda n} \sum_{i=1}^n \psi_i(\hat{y}_i^{(t)})$

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

$p(w, \xi^*(w))$ $\quad \max_{\tilde{y}} l_i(\tilde{y}) - w^T \psi_i(\tilde{y})$

$\frac{\partial p}{\partial w^{(t)}} = \lambda w^{(t)} - \frac{1}{n}\sum_{i=1}^n \psi_i(\hat{y}_i^{(t)})$

a subgradient of p at $w^{(t)}$

a subgradient step with step-size β

$w^{(t+1)} = (1 - \beta\lambda) w^{(t)} + \beta \frac{1}{\lambda n}\sum_{i=1}^n \psi_i(\hat{y}_i^{(t)})$

thus batch FW on dual with step-size $\gamma = \beta\lambda$ is equivalent to batch subgradient on primal with stepsize $\beta = \frac{\gamma}{\lambda}$

a disadvantage of FW is can use line search (adaptive step-size subgradient method)

canonical FW step-size $\gamma_t = \frac{2}{t+2} \Rightarrow \beta_t = \frac{1}{\lambda}\frac{2}{t+2}$

λ - strong convexity parameter of primal

step-size in lecture 5?

note: subgradient convergence proof for μ-strongly convex obj.

needed $\beta_t \le \frac{1}{\mu} O(\frac{1}{t})$

$R \triangleq \max_{i,y} \|\psi_i(y)\|_2$

$p(\hat{w}^{(t)}) - p(w^*) \le \frac{8R^2}{\lambda(t+2)}$

$\hat{w}^{(t)} \triangleq \sum_{s=0}^t \beta_s^t w^{(s)}$ (weighted average)

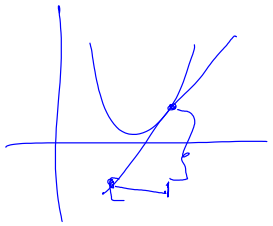
$\beta_s^t \propto s$

$\min_{s \le t} p(w^{(s)}) \le \sum_{s=0}^t \beta_s^t p(w^{(s)})$ for any convex combo

convexity $\le p(\sum_{s=0}^t \beta_s^t w^{(s)})$

FW analysis: $d(\alpha^*) - d(\alpha^{(t)}) \leq \frac{2C_f}{t+2}$; turns out that $C_f = \frac{4R^2}{\lambda}$ (Later today)

Let's look at FW gap & $\langle -\nabla f(\alpha^{(t)}), s^{(t)} - \alpha^{(t)} \rangle = \frac{1}{n} \sum_{i=1}^{n} \left(H_i(\hat{y}_i^{(t)}; w^{(t)}) - \sum_{y \in Y_i} \alpha_i(y) H_i(y; w^{(t)}) \right)$



$= p(w(\alpha^{(t)})) - d(\alpha^{(t)})$

(Lagrangian gap of lecture 9?)

[note: not always the case; see e.g. CRF objective]

FW gaps

* one can show, that $\min_{s \leq t} g_s \leq 3 \cdot \frac{2C_f}{t+2}$

weighted average

it's also true that $g^{FW}(\hat{\alpha}^{(t)}) \leq 3 \frac{2C_f}{t+2}$ (optimization "error") when $g(\cdot)$ is convex

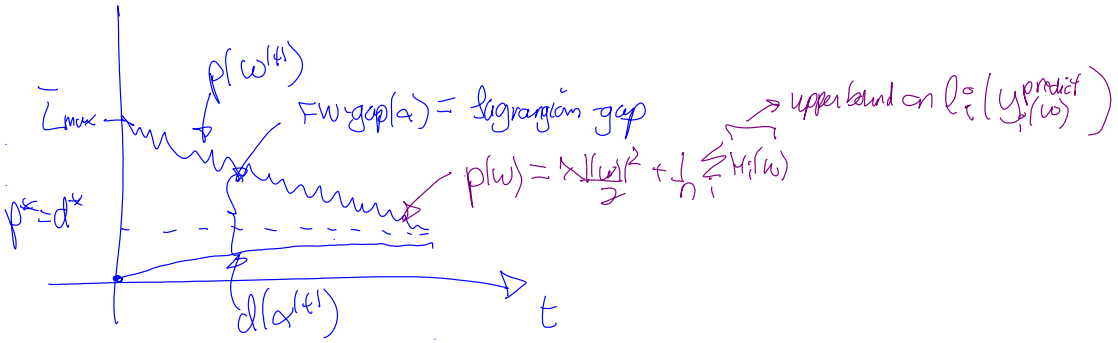
this is case for quadratic functions

↳ batch FW on dual for svmstruct $g^{FW}(\hat{\alpha}^{(t)})$

also gives $p(\hat{w}^{(t)}) - p(w^*) = O(\frac{1}{t})$

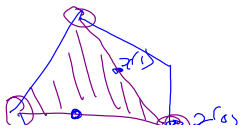
$p(0) = \frac{1}{n} \sum_{i=1}^{n} \max_{\tilde{y}} l_i(\tilde{y}) = \bar{L}_{max}$

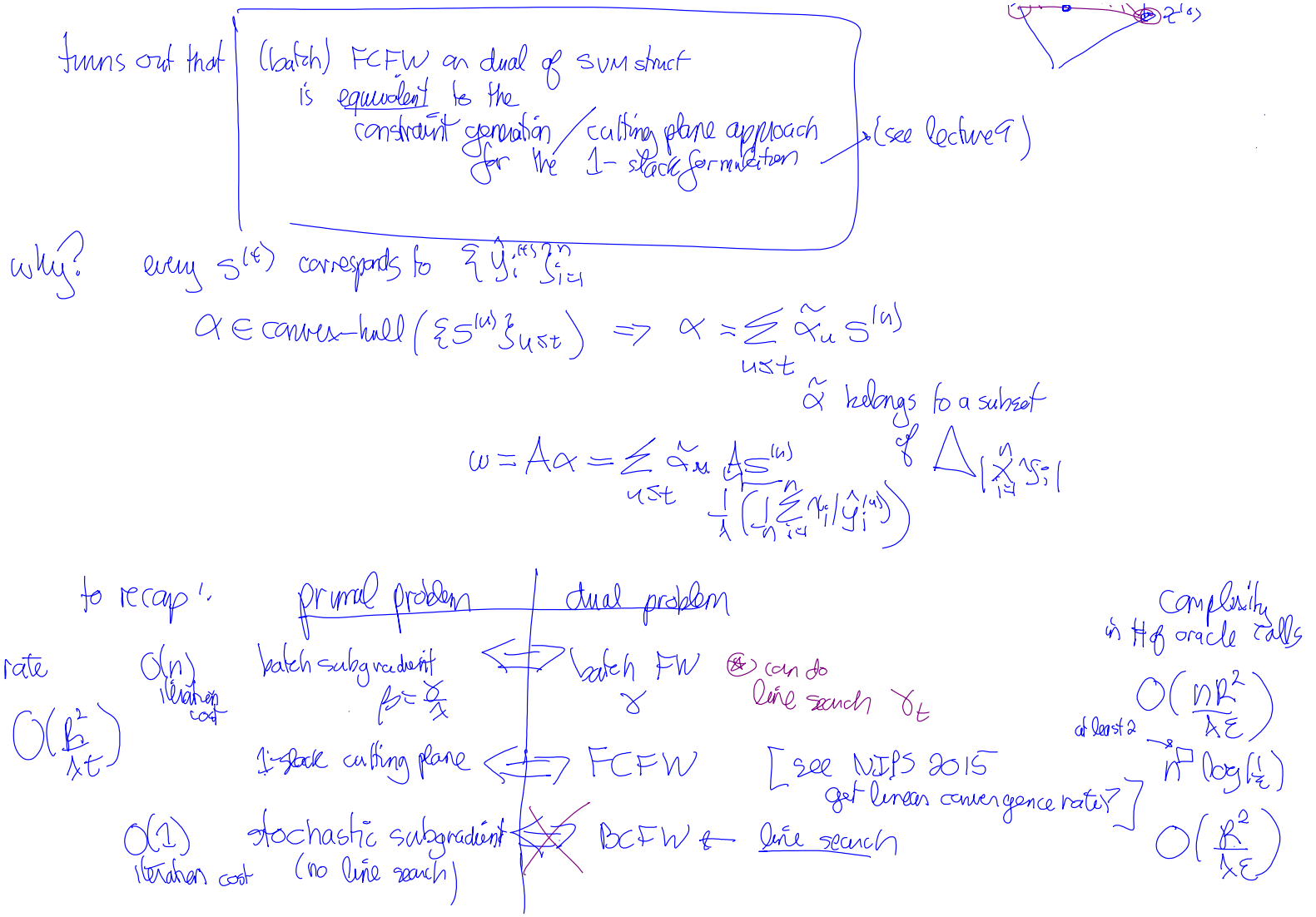
$d(\alpha^{(0)}) = 0$



another variant of FW: FCFW "fully-corrective FW"

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





pointers:

- the usual pointer for FW for structured SVM (as in previous lecture):
  - Block-Coordinate Frank-Wolfe Optimization for Structural SVMs, S. Lacoste-Julien, M. Jaggi, M. Schmidt and P. Pletscher, ICML 2013
    http://jmlr.csail.mit.edu/proceedings/papers/v28/lacoste-julien13-supp.pdf

- the linear convergence of FCFW is given in:
  On the Global Linear Convergence of Frank-Wolfe Optimization Variants, S. Lacoste-Julien and M. Jaggi, NIPS 2015
  http://arxiv.org/abs/1511.05932 | NIPS link

(already cited in last lectures)

- a good application of using FCFW when the linear oracle is expensive is given in this paper:
  Barrier Frank-Wolfe for Marginal Inference, R. Krishnan, S. Lacoste-Julien and D. Sontag, NIPS 2015
  https://arxiv.org/abs/1511.02124

- the generalization of the observation that Frank-Wolfe optimization sometimes reduced to the subgradient method on the primal is given in:
  - F. Bach. Duality between subgradient and conditional gradient methods.
    SIAM Journal of Optimization, 25(1):115-129, 2015
    http://www.di.ens.fr/~fbach/sg_cg_fbach_siopt.pdf