Lecture 16 - FW

Thursday, March 12, 2020 13:35

Away-step FW

**Comment:**\( \alpha_k = \sum_u \frac{d_j}{\|d_u\|^2} s_u \)

\[ x_{kn} = (1-\delta_k) x_k + \delta_k s_k \]

\[ = \sum_u (1-\delta_k) s_u + \delta_k s_k \]

**FW step moves most uniformly away from active set**

**unless step size** \( \delta_k = 1 \), FW never removes a corner from its expansion

**Zig-zag phenomenon close to boundary**

\[ - \nabla f(x_k) > 0 \Rightarrow \frac{\|d_k\|}{\|\nabla f(x_k)\|} \rightarrow 0 \Rightarrow \frac{\delta_k}{\|d_k\|} \rightarrow 0 \]

**Slow rate**

**Away-step FW fix:** (solves Zig-Zagging problem)

in addition to compute:\( s_k = \arg \max_{s} \langle \nabla f(x_k), s \rangle \)

also compute: \( \lambda_k = \arg \min_{\lambda} \langle \nabla f(x_k), \lambda \rangle \)

**Sublevel set of a strongly convex**

when \( \lambda_k \) in the relative interior of \( \Lambda \)

**Strongly convex**

\[ O(\lambda_k^{-p}) \]

For strongly convex, \( p \)

IFT6132 Page 1
\[ V_e = \arg \max_{v \in \text{achieved}(x_e)} \langle d(y), s \rangle \]

\[ \text{AFW} = S_e - x_e \]
\[ d_A = x_e - v_e \]

* AFW picks the direction with best inner product.

\[ \text{AFW} = \arg \max \langle d_A, d_f(x_e) \rangle > \langle d_{FW}, d_f(x_e) \rangle \]

\[ \text{AFW} \]

\[ \text{can be seen when } d_f(x_e) \text{ is a } i \text{-to-achieve } \]

* If use \( d_A \), let \( x_f = \arg \min_{\gamma \in [0, \gamma_{\max}]} \)

\[ x_f = \frac{d_A}{\gamma} \]

\[ x_{f+1} = x_e + \gamma (x_e - x_e) \]

\[ = \frac{(1 + \gamma_e) \alpha u \gamma_e}{\gamma} - \gamma e V_e \]

\[ \lambda \text{ or be coefficient of } \gamma_e \]

\[ (1 + \gamma_{\max}) \gamma_e - \gamma_{\max} = 0 \]

\[ \Rightarrow \gamma_{\max} = \frac{\alpha}{1 - \alpha} \]

* When \( \gamma_e = \gamma_{\max} \)

"Drop step" \rightarrow remove \( v_e \) from expansion

* When run AFW:

either you maintain some expansion of \( x_e = \frac{\alpha}{\gamma_{\max}} u \)

or

you have a feasibility oracle + away-step oracle

[See NIPS 2016 paper]

By Mosh 
Carlen

\[ \text{assumption to prove convergence} \]

\[ x_e \]
AFW has linear convergence rate on polytopes when $f$ is strongly convex.

\[ \text{Convex combination of Fw, SDA} \rightarrow \text{pairwise Fw} \]

\[ c_{\text{Convex}} + c_A = \frac{1}{2} z - \frac{1}{2} z_t - v_t = \frac{1}{2} z - v_0 \]

\[ \langle -D(f(x)), g_D \rangle = g_{\text{Fw}} + g_A \]

Optimizing Fw: $\theta_0 = \frac{1}{2} D(f(x)), \theta_0 \geq \max \{ e_{\text{Fw}}, g_A \}$

\[ \theta_0 \geq \theta_{\text{Fw}} \]

Note: if $M = \text{conv}(A)$ where $A$ is some finite set (called "atoms")

\[ \text{LMO}(r) : \min \langle s, r \rangle = \max \langle a, r \rangle \quad \text{s.t.} \quad A \subseteq M \]

\[ \text{LMO} \rightarrow \text{min cost network flow} \]

e.g. $A \rightarrow$ integer flows $\text{conv}(A) \rightarrow$ flow polytope

$a \rightarrow$ clique assignment in graph $\text{conv}(A) \rightarrow$ matching polytope

More properties of Fw:

2) Convergence for non-convex $f$:

Aside: necessary first order condition for constrained opt.

\[ \min_{x \in M} f(x) \quad x^* \text{ is a local min} \]

\[ \Rightarrow \langle D(f(x^*)), s - x^* \rangle > 0 \quad \forall s \in M \]
min \{ f(x) \mid x \in M \}

\[ x^* \text{ is a local min} \]

\[ \Rightarrow < Df(x^*), 1 - x^* > \geq 0 \quad \forall s \in M \]

\[ \Leftarrow \min_{s \in M} < Df(x^*), s - x^* > \geq 0 \]

\[ \Leftrightarrow \max_{s \in M} < -Df(x^*), s - x^* > \leq 0 \]

\[ \text{FW-gap}(x^*) \]

\[ \text{quantify the "non-stationarity"!} \]

\[ \text{see L.J. 2016 and iv} \]

\[ \min_{s \in M} \text{gap}(rs) \leq O\left( \frac{1}{\sqrt{t}} \right) \]

\[ \text{for FW with the same set} \]

\[ \text{"non-convex" FW} \]

\[ f \text{ is } \text{L-smooth and bounded} \]

\[ M \text{ is bounded (convex) but } f \text{ is not necessarily convex} \]

4) affine covariance of FW

let \( \tilde{M} \) be a new domain s.t. \( \tilde{M} \rightarrow M \) i.e. \( M = A\tilde{M} \)

\[ \tilde{f}(\tilde{x}) := f(A\tilde{x}) \]

\[ \min_{\tilde{x} \in \tilde{M}} \tilde{f}(\tilde{x}) = \min_{x \in M} f(Ax) = \min_{x \in M} f(x) \]

affine covariance of FW:

- If run FW on \( \tilde{f} \) it \( \tilde{x}_t \) to get \( \tilde{x}_t \) it\( \tilde{a}_{t} \)

- Then \( x_t := A\tilde{x}_t \) corresponds to running FW on \( f \) in \( M \)

\[ \text{modulo tie-breaking} \]

why? inner product is affine invariant

\[ < D_{\tilde{x}} \tilde{f}(\tilde{x}), \tilde{z} > = < D_{x} f(Ax), Ax \cdot \tilde{z} > \]

\[ \tilde{x}_t = \text{argmin}_{\tilde{x} \in \tilde{M}} < \tilde{z}, D_{\tilde{x}} \tilde{f}(\tilde{x}) > \quad \tilde{x}_t \in A\tilde{x}_t \]

\[ < \tilde{z}, A^T D_{x} f(x) > \]

\[ \tilde{x}_t = \text{argmin}_{\tilde{x} \in \tilde{M}} < \tilde{z}, D_{\tilde{x}} \tilde{f}(\tilde{x}) > \quad \tilde{x}_t \in A\tilde{x}_t \]

\[ s_t = \text{argmin}_{s \in \mathbb{R}^d} < s, D_{x} f(x) > \quad s_t \in A\tilde{s}_t \]

\[ s_t \rightarrow A\tilde{s}_t \quad (\text{modulo tie-breaking}) \]
\[ \Rightarrow \text{we want affine invariant metrics} \Rightarrow C_\gamma \leq L_{\text{aff}} \text{dist}_B(\gamma)^2 \]