Today: finish prox example
- catalyst → accelerate
  - non-convex opt
  - submodular opt

finish prox example

To be useful, need $\text{prox}_g^\alpha$ to be efficiently computable

$$
\text{prox}_g^\alpha(z) = \underset{w}{\arg\min} \|w\|_1 + \frac{1}{2\alpha} \|w - z\|_2^2
$$

"soft-thresholding" \(\frac{\text{sgn}(z_i) [\alpha i - 8]}{\text{component-wise}} \) if \( |z_i| \geq \alpha \)
0 o.w.

used edge for lasso: $\ell_1$-reg-least-square

FISTA → accelerated prox. gradient method
FISTA for batch lasso

* scikit-learn → use SAGA for lasso $\ell_1$-reg

\[
\text{prox}_{\alpha g_A} : w_{k+1} = \text{prox}_g^\alpha \left( w_k - \alpha \left[ \sum_{i=1}^n \text{sgn}(w_i) - g \right]_{\text{soft}} + \frac{1}{n} \sum_{i=1}^n g_i \right)
\]

Could accelerate using "Catalyst"

Catalyst algorithm [Lin, Mairal, & Harchaoui, NIPS 2015]

"meta-algorithm": outer loop which uses a linearly convergent algo in inner loop to get overall acceleration (\?)

main idea: use the accelerated proximal point algorithm

proximal pt. algo: is proximal gradient with $\alpha = 0$

\[
W_{k+1} = \text{prox}_g^\alpha (W_k)
\]
(to solve $\min_w \Omega(w)$)
catalyst alg. \( (\text{for } \mu\text{-strongly convex } F(w)) \)

\[
\text{let } q = \frac{\mu}{\mu + 1} \quad (\times \text{ is algorithmic parameter})
\]

repeat:

\[
\begin{align*}
\hat{w}_{t+1} & \leftarrow \text{argmin} \ F(w) + \frac{1}{2\eta} \|w - z_t\|^2 \quad \text{s.t. } G_{\delta}(w) - \min \ G_{\delta}(w) \geq 2\epsilon_t \\
& \text{using inner loop optimization alg., } \text{e.g. SAGA or AFW} \\
& \text{accelerated Nesterov' trick piece} \\
z_{t+1} = w_{t+1} + \beta_{t+1} (w_{t+1} - w_t)
\end{align*}
\]

\( \beta_{t+1} \text{ is found using fancy equations so that everything works} \)

\[
\beta_{t+1} = \frac{\alpha_t (1 - \alpha_t)}{\alpha_t^2 + \alpha_{t+1}} \
\text{(pick } \alpha_{t+1} \in [0, 1] \text{)}
\]

Catalyst trick: use \( \alpha \in \mathbb{E} \)

\( \text{s.t. overall # of inner loop calls} \)

\( \text{give overall acceleration} \)

\( \text{with clever analysis of inner starting} \)

acceleration results:

if inner loop alg. has convergence \( \exp(-\frac{\mu}{\xi} t) \) \( \tilde{w} \geq \mu + 1 \)

then with correct constants

\[
(\mu\text{-strongly convex } F) \text{ linear rate: } \rho = \frac{1}{1 + \frac{\mu}{\xi}} \quad \tilde{w} \geq \mu + 1
\]

\[
(F \text{ convex case}) \quad \frac{1}{t} \text{ or } F \text{ changes } \frac{1}{t^2}
\]

results can get \( \text{theory) accelerated SAGA, AFW etc.} \)
**Non-convex optimization**

Recall: FW with line search on $f$ non-convex

$$\min_{w \in \mathbb{R}^d} g(w) \leq O\left(\frac{1}{\sqrt{B}}\right)$$

Convex: $E[f(w_t)] - f^* \leq 3$

Non-convex: $\sqrt{E[\|f(w_t)\|^2]} \leq 3$

Gradient method:

$$f(w_t) \leq f(w_c) + \frac{5}{2} \|w-w_c\|^2 + \frac{1}{2} \|w_t-w_c\|^2$$

$$w_{t+1} = w_t - \frac{1}{L} Df(w_t)$$

$$\Rightarrow f(w_{t+1}) \leq f(w_c) - \frac{1}{2L} \|Df(w_t)\|^2$$

---

**Faster nonconvex optimization via VR**

*(Reddi, Hefny, Sra, Poczos, Smola, 2016; Reddi et al., 2016)*

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Nonconvex (Lipschitz smooth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>$O\left(\frac{1}{\sqrt{L}}\right)$</td>
</tr>
<tr>
<td>GD</td>
<td>$O\left(\frac{2}{L}\right)$</td>
</tr>
<tr>
<td>SVRG</td>
<td>$O\left(n + \frac{n^{2/3}}{\epsilon}\right)$</td>
</tr>
<tr>
<td>SAGA</td>
<td>$O\left(n + \frac{n^{2/3}}{\epsilon}\right)$</td>
</tr>
<tr>
<td>MSVRG</td>
<td>$O\left(\min\left(\frac{1}{L}, \frac{n^{2/3}}{\epsilon}\right)\right)$</td>
</tr>
</tbody>
</table>

**Remarks**

New results for convex case too; additional nonconvex results
For related results, see also *(Allen-Zhu, Hazan, 2016)*

**Linear rates for nonconvex problems**

$$\min_{\theta \in \mathbb{R}^d} g(\theta) = \frac{1}{n} \sum_{t=1}^{n} f_t(\theta)$$

The **Polyak-Łojasiewicz (PL) class of functions**

$$g(\theta) - g(\theta^*) \leq \frac{1}{2\mu} \|\nabla g(\theta)\|^2$$

*(Polyak, 1963); (Łojasiewicz, 1963)*
Linear rates for nonconvex problems

\[ g(\theta) - g(\theta^*) \leq \frac{1}{2\mu} \| \nabla g(\theta) \|^2 \quad \text{or} \quad \mathbb{E}[g(\theta_t) - g^*] \leq \epsilon \]

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Nonconvex</th>
<th>Nonconvex-PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>( O\left( \frac{1}{\epsilon^2} \right) )</td>
<td>( O\left( \frac{1}{\epsilon} \right) )</td>
</tr>
<tr>
<td>GD</td>
<td>( O\left( \frac{n}{\epsilon} \right) )</td>
<td>( O\left( \frac{n}{2\mu \log \frac{1}{\epsilon}} \right) )</td>
</tr>
<tr>
<td>SVRG</td>
<td>( O\left( n + \frac{n^{2/3}}{\epsilon} \right) )</td>
<td>( O\left( \left( n + \frac{n^{2/3}}{2\mu} \right) \log \frac{1}{\epsilon} \right) )</td>
</tr>
<tr>
<td>SAGA</td>
<td>( O\left( n + \frac{n^{2/3}}{\epsilon} \right) )</td>
<td>( O\left( \left( n + \frac{n^{2/3}}{2\mu} \right) \log \frac{1}{\epsilon} \right) )</td>
</tr>
<tr>
<td>MSVRG</td>
<td>( O\left( \min \left( \frac{1}{\epsilon}, \frac{n^{2/3}}{\epsilon} \right) \right) )</td>
<td>_</td>
</tr>
</tbody>
</table>

Variant of nc-SVRG attains this fast convergence!

(Reddi, Hejny, Sra, Poczos, Smola, 2016; Reddi et al., 2016) 22

---

Submodular optimization

Submodularity is an analog of convexity for functions,

For set functions (combinatorial opt.)

\( F : 2^V \to \mathbb{R} \)

\( V = \{ 1, \ldots, d \} \) is "ground set"

\( 2^V = \{ V \to \mathbb{R} \} \) = set of all subsets of \( V \)

Concrete example:

Fixing model \( y_i \in \{ 0, 1 \} \)

\( E(y) = \sum_i y_i - \sum_{i,j} y_i y_j \quad \text{(mixture)} \)

When \( E(y) > 0 \)

"attractive potential"

\( E(y) \) is submodular

MAF here is called "Associative Markov network" "AMN"

\( F(A_y) \) where \( A_y = \sum_i y_i = 4 \)

Can minimize \( E(y) \) by using "graph cut" alg.

(\text{or } F(A_y))
\[ F \text{ is submodular } \iff F(A) + F(B) \geq F(AB) + F(A \cup B) \forall A B \]

\[ \iff \text{function } A \mapsto F(A \cup B) - F(A) \text{ is non-increasing for } \emptyset \neq A \subsetneq B \]

\[ \text{1. } F(A \cup B) - F(A) \leq F(B) - F(A) \]

"diminishing return property"

\[ \implies \text{intuitively, that greedy alg are not "too bad" for maximization} \]

\[ \star F(A) = g(\frac{1}{n}) \quad \text{if } g \text{ is concave than } F \text{ is submodular} \]

\[ \star \text{link with convexity: Lovasz extension (cts. fed.)} \]

\[ \star \text{embedded sets as corners of hypercube in dimension } \mathbb{R}^n = \bigvee_{A \subseteq \mathcal{V}} \]

\[ \text{Lovasz extension } f \text{ extends } F(A) \text{ from corners to entire hypercube using convex interpolation} \]

\[ f(w) = F(A) \quad \text{when } w = v(A) \]

\[ F \text{ is submodular } \iff \text{Lovasz extension is convex} \]

\[ \star \text{can write } f(w) = \max_{s \in \mathcal{V}} \langle s, w \rangle \quad \text{-- this can be computed efficiently using greedy alg} \]

"base poly" \[ \min_{A \subseteq \mathcal{V}} F(A) = \min_{w \in \bigvee_{A \subseteq \mathcal{V}}} \left( \max_{s \in \mathcal{V}} \langle s, w \rangle \right) \]

\[ \iff \min_{w \in \bigvee_{A \subseteq \mathcal{V}}} \frac{f(w)}{\|w\|} \]

\[ \star \text{with } l_2 \text{-regularization, use duality to get } \]

\[ \text{smooth alg, } \min_{s \in \bigvee_{A \subseteq \mathcal{V}}} \frac{1}{2} \|s\|^2 \]

\[ \implies \text{use "mirror prox alg" } \}

\[ \text{varient FCFW alg for submodular qts.} \]