Today is optimization (of $f(w)$)

$$S(w) = R(w) + \frac{1}{n} \sum_{i=1}^{n} S(x^{(i)}, y^{(i)}, w)$$

Structured $SOM$: $R(w) = \frac{1}{2} \|w\|_2^2$ \quad $S(x^{(i)}, y^{(i)}, w) = \max_{\tilde{y} \in \mathbb{R}(x^{(i)})} \left( R(w, \phi(x^{(i)}, \tilde{y}) + l(y^{(i)}, \tilde{y})) - <w, \phi(x^{(i)}, y^{(i)})> \right)$

(structured hinge loss)

let $l_i(y) \equiv l(y^{(i)}, \tilde{y})$ \quad $y_i \equiv y(x^{(i)})$

$H_i(w) \equiv S(x^{(i)}, y^{(i)}, w) = \max_{\tilde{y} \in \mathbb{R}(x^{(i)})} \left( l_i(y) - <w, \phi(x^{(i)}, \tilde{y})> \right)$

(structured SVM objective, non-smooth unconstrained form)

$$\max_{w} \left( \|w\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} H_i(w) \right)$$

$$\max_{y \in \mathbb{R}} \left( l_i(y) - <w, \phi(x^{(i)}, y)> \right)$$

$|x| = \max x - x$
Convex analysis recap

\[ |x| = \max \{ x, -x \} \]

\[ f \text{ is convex} \quad \text{iff} \quad f(px + (1-p)y) \leq pf(x) + (1-p)f(y) \]

\[ p \in [0, 1] \]

subdifferential of \( f \) at \( x \)

subgradient: \( v \in \partial f(x) \), “\( v \) is a subgradient of \( f \) at \( x \)”

\[ \iff \forall y \in \text{dom} f \quad f(y) \geq f(x) + \langle v, y-x \rangle \]

Standard assumptions on \( f \):

Smooth:

\( \nabla f \) is 1-Lipschitz continuous i.e. \( \| \nabla f(x) - \nabla f(y) \| \leq L \| x-y \| \)

Hessian

\[ L = \max_{x \in \text{dom} f} \| H(x) \| \]

Non-smooth:

\( \partial f \) is bounded i.e. \( \| v \| \leq B \) for all subgradients \( v \) of \( f \)

Strong convexity of \( f \):

\( f \) is strongly convex \( \iff \quad f(x) \geq f(y) + \langle \nabla f(y), x-y \rangle + \frac{1}{2} \| x-y \|^2 \)
$f$ is $\mu$-strongly convex $\iff$ $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\mu}{2} \|y-x\|^2$ by (convex)

$\forall y, x$ for any $x \in \text{dom} f$

- If $f$ is $C^2$, $\mu = \min_{x \in \text{Hessian}(f)} I$

Properties:
- $f$ $\ell$-Lipschitz $\implies$ $f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\ell}{2} \|y-x\|^2$
- $f$ $\mu$-strongly convex $\implies$ $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\mu}{2} \|y-x\|^2$

Question: Is $f(x)$ is bounded and convex,

$\implies$ $f$ is $\ell$-Lipschitz

Answer: No; counterexample: $f(x) = -\sqrt{x}$ on $[0,1]$

(f is not Lipschitz continuous)

Landscape of convergence rates: $\rightarrow$ for first order methods

- Convergence result: $f(x_k) - f(x^*) \leq \frac{\text{const.}}{\sqrt{k}}$ (non-smooth rate)
- Function smoothness $\rightarrow$ $\rightarrow$
- Suppose $\text{dist}(x_k, x^*) \leq s_0$

Assumptions | Rate deterministic | Rate stochastic
--- | --- | ---
- Non-smooth | $O\left(\frac{B^2}{\sqrt{k}}\right)$ | $O\left(\frac{1}{\sqrt{k}}\right)$
- 1) | $O\left(\frac{1}{\sqrt{k}}\right)$ | $O\left(\frac{1}{\sqrt{k}}\right)$

Stochastic method
<table>
<thead>
<tr>
<th>Smooth L-Lipschitz</th>
<th>$O\left(\frac{L}{K}\right)$</th>
<th>$O\left(\frac{L^2}{K}\right)$</th>
<th>$O\left(\frac{L}{\sqrt{K}}\right)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{10}$</td>
<td>Gradient method</td>
<td>Nesterov method</td>
<td>$O\left(\frac{1}{\sqrt{K}}\right)$</td>
</tr>
</tbody>
</table>

$\bar{\mathcal{R}}$ is $\mu$ strongly convex

\[ O\left(\frac{L^2}{K}\right) \text{ subgradient method} \]

\[ O\left(\exp\left(-\frac{L^2}{2K}\right)\right) \text{ gradient method} \]

\[ O\left(\frac{L^2}{\sqrt{K}}\right) \text{ Nesterov method} \]

\[ O\left(\frac{L}{K}\right) \rightarrow E_{x\sim P} f(x, z) \]

\[ \sum_{i=1}^{\bar{\mathcal{R}}} \mathbb{E}_{z\sim S_i} f_i(x) \]

\[ \text{true expectation} \]

\[ \text{SGD} \]

\[ \frac{L^2}{K} \text{ stochastic averaged gradient method} \]

\[ \text{using SAM, etc.} \]

\[ \text{Note: projection gives same rate} \rightarrow \text{reason is contraction property of projection} \]

\[ \|P_C(x) - P_C(x^*)\| \leq \frac{1}{10} \|

\[ x_{t+1} = x_t - \frac{1}{L} \nabla f(x_t) \]

\[ \text{Projected gradient method for min } f(x) : x_{t+1} = P_C \left[ x_t - \frac{1}{L} \nabla f(x_t) \right] \]

\[ x \in C \]

\[ \text{Euclidean projection on } C \text{ i.e.} \]

\[ P_C(y) = \arg\min_{x \in C} \|x - y\|^2 \]

\[ \min_{x \in C} \]

\[ \|P_C(y) - P_C(x^*)\| \leq \frac{1}{10} \|y - x^*\| \]

\[ \text{unconstrained} \]

\[ \Rightarrow \text{so rates transfer to "projected" versions} \]
Complexity in optimization

"work" to get to ε-suboptimality $O\left(\frac{1}{\varepsilon^2}\right)$ rate $\Rightarrow O\left(\frac{1}{\varepsilon^2}\right)$ # of iterations

If want to compare methods, include also "cost" of iterations

Stochastic subgradient method

Say want to solve $\min_{x \in C} f(x)$ with $f(x) = \frac{1}{2} \sum_{i=1}^{n} h_i(x, \xi_i)$

Assumptions:
1) Projection on $C$ is cheap
2) $f$ is convex in $x$
3) We have stochastic oracle which gives $\mathbf{G}_t$ at time $t$, a random direction such that $\mathbb{E}[\mathbf{G}_t \mid x_t] = \nabla f(x_t)$

$\mathbb{E}[\mathbf{G}_t \mid "rand" ]$

E.g. if $f$ is differentiable in $x$ and "well behaved"

$\mathbf{G}_t(x, \xi_t)$ for $\xi_t \sim \mathcal{D}$ then $\mathbb{E}[\mathbf{G}_t(x, \xi_t)] = \nabla f(x)$

More specifically if $f(x) = \sum_{i=1}^{n} \xi_i h_i(x)$

Then pick it with at random $i \in \{1, \ldots, n\}$
and letting $g_t = S_t(x_t)$

$\implies \mathbb{E} g_t | x_t = \sum_{i=1}^{n} \mathbb{E} s_i (x_t) = s(x_t)$

4) $\mathbb{E} \| g_t \|^2 \leq B^2$ [finite variance condition]

$\mathbb{E} \| f_t(x) \| \leq B$ is sufficient for this then

\[ \text{algorithm:} \]
\[ x_0 \in C \text{ initialization} \]
\[ \text{for } t = 0, \ldots, T-1 \]
\[ \text{get } g_t \text{ from oracle, step-size} \]
\[ \text{let } x_{tn} = \varphi_{C_t} [ x_t - \delta_t g_t ] \]
\[ \text{output } x_T = \frac{1}{T} \sum_{t=0}^{T} x_t \text{ where } \delta_t \propto t \]

"weighted average"

$\delta_t = \frac{2t}{T(T+1)}$

pointers:
- book for rates of convergence & lower bound:
  - Nesterov, "Introductory Lectures on Convex Optimization", 2004
- proof for convergence of weighted average stochastic subgradient method:
  