(Interlude on computer: show memory/[illegible] and gradient Test.)

une 11 '18

Lecture 3: Modern 1st-Order Methods For Structured Problems

▷ GRADIENT DESCENT

$$\boxed{\min_x f(x), \quad \text{assuming } f \text{ is } \mathbb{R}\text{-valued}, \; f \in \Gamma_0(\mathbb{R}^n), \; \nabla f \text{ } L\text{-Lipschitz}}$$ — e.g., $0 \preceq \nabla^2 f(x) \preceq L \cdot I$

$$x_{k+1} = \operatorname*{argmin}_x f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \underbrace{\frac{L}{2}\|x - x_k\|^2}_{\text{discuss, vis-a-vis}}$$

$$= x_k - \frac{1}{L}\nabla f(x_k)$$

$$= x_k - t\nabla f(x_k) \text{ w/ stepsize/learning-rate } t = 1/L$$

Cond. Gra / Frank-Wolfe

Newton

See Bubeck

is a

nt method.

cent linesearch

t help

ntically

worst case)

<u>Convergence Analysis</u> (Vandenberghe's notes, i.e., Nesterov's book), for $t = \frac{1}{L}$

Let x^* be any optimal sol'n

$$f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \quad \text{"Descent Lemma"}$$

$$= f(x_k) + \langle \nabla f(x_k), -\frac{t}{L}\cdot\nabla f(x_k) \rangle + \frac{L}{2}\|\frac{-1}{L}\nabla f(x_k)\|^2$$

See picture

$$= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \quad (\Rightarrow \text{descent method}) \quad f(x_{k+1}) \le f(x_k)$$

$$\le \overbrace{f(x^*) + \langle \nabla f(x_k), x_k - x^* \rangle} - \frac{1}{2L}\|\nabla f(x_k)\|^2 \quad \text{via convexity.}$$

$$= f(x^*) + \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2 \right)$$

$$= f(x^*) + \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right).$$

Add, for $i = 1, \dots, k$,

x) - f(·) ≤

$$\frac{1}{k}\sum_{i=1}^{k}\left(f(x_i) - f(x^*) \right) \le \frac{1}{k}\frac{L}{2}\sum_{i=1}^{k}\left(\|x_{i-1} - x^*\|^2 - \|x_i - x^*\|^2 \right) \quad \text{telescopes}$$

$$= \frac{1}{k}\frac{L}{2}\left(\|x_0 - x^*\|^2 - \underbrace{\|x_k - x^*\|^2}_{\ge 0} \right)$$

$$\le \frac{1}{k}\frac{L}{2}\|x_0 - x^*\|^2$$

so ~~f~~ and, since it was a descent method, $f(x_k) \le f(x_i) \ \forall i = 1, \dots, k$

so $$\boxed{f(x_k) - f^* \le \frac{L}{2k}\|x_0 - x^*\|^2} < \varepsilon$$

or, if we want $f(x_k) - f^* < \varepsilon$, take $k > \frac{L}{2}\|x_0 - x^*\|^2 \cdot \frac{1}{\varepsilon}$, i.e., $\boxed{O(\frac{1}{\varepsilon}) \text{ iterations}}$

"sub-linear"

* ~~Short~~ $\boxed{\text{Asymptotic Worst-Case Result Only!}}$ Discuss...

Is this rate good? Tight?

$x_{k+1} \in \text{span}\{x_0, \nabla f(x_0), \ldots, \nabla f(x_k)\}$

Thm: Nestrov 1980s, 2003 book:

No 1st order method can ~~beat~~ always guarantee

$$f(x_k) - f(x^*) < \tfrac{3}{32} \cdot L \cdot \frac{\|x_0 - x^*\|^2}{k^2} \quad \text{for } k \le \tfrac{1}{2}(n-1).$$

Suppose f is strongly cvx

$$\forall x, \quad \forall y, \quad f(y) \ge \underbrace{f(x) + \langle \nabla f(x), y - x\rangle + \mu/2 \, \|y-x\|^2}_{} \qquad \text{"sc"}$$

$$\min \phi(y) \ge \min \psi(y)$$

$$f(y) \ge g(y) \implies \min_y f(y) \ge \min_y g(y) \quad \nwarrow \; y = x - \tfrac{1}{\mu}\nabla f(x)$$

So

$$\min_y f(y) = f(x^*) \ge f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2 = \min_y g(y)$$

PL Polyak-Łojasiewicz Ineq: $\forall x, \quad \tfrac{1}{2}\|\nabla f(x)\|^2 \ge \mu \cdot (f(x) - f^*)$

weaker than sc (ie, ~~don't~~ unique sol'n, while sc does)

"essential sc", "restricted sc", "error bnd condition"... $\Rightarrow$ PL.

[Can bound sub-optimality by gradient]

Thm (Karimi, Nutini, Schmidt '16)

Grad. Descent w/ $t = \frac{1}{L}$, ∇f L-Lipschitz, and μ-PL, satisfies

$$f(x_k) - f^* \le \left(1 - \tfrac{\mu}{L}\right)^k \cdot (f(x_0) - f^*)$$

proof:

as before, $f(x_{k+1}) - f(x_k) \le \frac{-1}{2L}\|\nabla f(x_k)\|^2$

$\le -\frac{\mu}{L}(f(x_k) - f^*)$ via PL

re-arrange, subtract f^*,

$$f(x_{k+1}) - f^* \le (1 - \mu/L)(f(x_k) - f^*). \quad \blacksquare$$

Rates to reach ε-sol'n		to go from $\varepsilon_0 = 1$ to $\varepsilon' = 10^{-2} \cdot \varepsilon_0$, need __ more it.
subgradient desc.	$O(1/\varepsilon^2)$	10,000 x more
gradient desc.	$O(1/\varepsilon)$	100 x more
accel. gra. desc.	$O(1/\sqrt{\varepsilon})$	10 x more
linear (eg., sc)	$O(-\log(\varepsilon))$	~~about 2 2 more~~ 2·constant more
quadratic (eg, Newton)	$O(\log(-\log(\varepsilon)))$	1 more

these depend on ε_0 (eg, $\varepsilon_0 = 1$)

▷ NESTEROV'S ACCELERATED METHOD – "optimal"

Extends heavy-ball method (analysis for quadratics in, eg, Bertsekas)

"momentum" $x_{k+1} = x_k - t_k \nabla f(x_k) + s_k(x_k - x_{k-1})$.

See

distill.pub/2017/momentum

Nesterov's Method, variant

~~$x_{k+1} = y_k - t\nabla f(y_k)$~~

~~$y_{k+1} = x_{k+1} +$~~

$$\boxed{\begin{aligned} x_k &= P\left[y_{k-1} - t\nabla f(y_{k-1})\right] \\ y_k &= x_k + \frac{k}{k+3}(x_k - x_{k-1}) \end{aligned}}$$

Thm For ∇f L-Lipschitz, if $t = \frac{1}{L}$, $f(x_k) - f^* = o(\frac{1}{k^2})$.

(see Vandenberghe, or my $\overset{4720}{\text{notes}}$ p55) for proof.

- Not a descent method!
- If f is strongly cvx, ought to know the constant μ in order to exploit.
- Extends to many variants, eg, proximal "FISTA" → discuss slow popularity
- Tricky to actually get (x_k) to converge, not just $(f(x_k))$ →

▷ NON-SMOOTH.

Suppose $f \in \Gamma(\mathbb{R}^n)$ but ∇f doesn't exist (∂f does)

Apply smooth method, hope for the best?

(p.40...) No, ex (Shor '98), even if you don't hit pts of non-diff., it messes up. (Wolfe), even if convex

Subgradient method

- $x_{k+1} = x_k - t_k \cdot d_k$ for any (deterministic) $d_k \in \partial f(x_k)$, ~~assuming~~
- $f \in \Gamma_0(\mathbb{R}^n)$, for simplicity assume also $\text{dom}(\partial f) = \mathbb{R}^n$
- Assume f (not ∇f) is Lipschitz continuous, constant L_0, ie., $\left(\|d_k\| \le L_0\right)$

A) Thm (8.13 in Beck) p.203

why? Not a descent method – not even descent direction

$$\min_{i\in 0,1,\dots,k} f(x_i^*) := f_{best}^k \le \frac{L_0 \cdot \text{dist}(x_0, \text{optimal})}{\sqrt{k+1}} \quad \text{i.e. } O\left(\frac{1}{\sqrt{k}}\right)$$

if $t_k = \dfrac{f(x_k) - f^*}{\|d_k\|^2}$ "Polyak's Stepsize Rule"

(skip mostly)

B) variant: <u>Thm (8.25 Beck)</u> If t_k isn't Polyak, but $\dfrac{\sum_{i=0}^{k} t_i^2}{\sum_{i=0}^{k} t_i} \to 0$ as $k \to \infty$,

then $f_{best}^{k} - f^* \to 0$ as $k \to \infty$

e.g., $t_k = \frac{1}{\sqrt{k+1}}$

c) variant: <u>Thm (8.28 Beck)</u> If $t_k = \dfrac{1}{\|g_k\| \cdot \sqrt{k+1}}$, $f_{best}^{k} - f^* = O\left(\dfrac{\log(k)}{\sqrt{k}} \right)$

and, <u>ergodic result</u>, if $\bar{x}_k = \dfrac{1}{\sum_{i=1}^{k} t_i} \sum_{i=1}^{k} t_i x_i$ is average, (note: can compute via a recursive update)

$f(\bar{x}_k) - f^* = O\left(\dfrac{\log(k)}{\sqrt{k}} \right)$

(remove log(k) if domain is compact)
(see eg. Bubeck)

D)
p. 265 <u>Thm (3.2 Bubeck)</u>

Assume we <u>project</u> onto C, radius is R, $\|g\| \le L_0$ again $(\forall g \in \partial f(x), \forall x)$

§3.5 Rates are tight. → Then if $t = \frac{R}{L_0} \cdot \frac{1}{\sqrt{k}}$, $f\Big(\underbrace{\frac{1}{k} \sum_{i=1}^{k} x_i}_{\bar{x}_k} \Big) - f^* \le \dfrac{R \cdot L_0}{\sqrt{k}}$

E) <u>Thm (8.31 Beck)</u> If f is also <u>μ-strongly convex</u>, take $t_k = \dfrac{2}{\mu(k+1)}$,

lso
3.9 in Bubeck

instead of $O\left(\frac{1}{\sqrt{k}}\right)$, and $f_{best}^{k} - f^* = O\left(\frac{1}{\mu k}\right)$, <u>instead of $O\left(\frac{1}{\sqrt{k}}\right)$</u>.

▷ <u>BETTER...</u>

Consider $\left(\min_x \; f(x) + g(x), \right)$ $f, g \in \Gamma_0(\mathbb{R}^n)$, ∇f is L-Lipschitz ($f = 0$ ok)
g isn't.

$\partial(f + g) = \nabla f + \partial g$, so subgradient descent is

$$x_{k+1} = x_k - t_k \cdot \big(\nabla f(x_k) + \overset{\text{for one}}{\partial g(x_k)} \big),$$

and as we saw, need $t_k \to 0$, so it's slow.

Instead, follow gra. desc.

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \Big(f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \tfrac{L}{2} \|x - x_k\|^2 \Big) + g(x) \Big)$$

$$= \underset{x}{\operatorname{argmin}} \; g(x) + \tfrac{L}{2} \| x - (x_k - \tfrac{1}{L} \nabla f(x_k)) \|^2 + \text{const.}$$

$$= \operatorname{prox}_{\frac{1}{L} g} \left(x_k - \tfrac{1}{L} \nabla f(x_k) \right).$$

PROXIMAL GRA. DESC.

Generalizes projected gradient descent,
no penalty from nonsmoothness of g (if you can compute prox).

recall, $x = \text{prox}_g(y)$ means $0 \in \partial g(x) + (x-y)$, i.e., $y \in (I + \partial g)(x)$

i.e.

$$\boxed{\text{prox}_g(y) = (I + \partial g)^{-1}(y)}$$

Other derivation:

$0 \in \nabla f(x) + \partial g(x)$ is optimality

$\iff 0 \in \frac{1}{L}\nabla f(x) + \frac{1}{L}\partial g(x)$

$\iff x - \frac{1}{L}\nabla f(x) \in x + \frac{1}{L}\partial g$

$\iff (I - \frac{1}{L}\nabla f)(x) \in (I + \frac{1}{L}\partial g)x$

$\iff \boxed{x = (I + \frac{1}{L}\partial g)^{-1}(I - \frac{1}{L}\nabla f)x}$ Fixed Pt. Eq'n. $x = Tx$

iterate $\qquad$ iterate $x_{k+1} = Tx_k$

$x_{k+1} = (I + \frac{1}{L}\partial g)^{-1}(I - \frac{1}{L}\nabla f)(x_k)$

i.e., $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) - \frac{1}{L}\partial g(x_{k+1})$ like implicit method, not explicit.

"SPECIAL CASE": $f \equiv 0$, then $x_{k+1} = (I + \frac{1}{\mu}\partial g)^{-1}x_k$ is the prox. pt. algorithm

$x_{k+1} = \text{argmin}_x\ g(x) + \mu/2\|x - x_k\|^2$

- General case is prox. gra. desc. aka "forward-backward"
- Extends to acceleration versions ("FISTA")
- No penalty on convergence rates compared to subgra. descent.
- Can you make a Newton version? Yes, but be careful!

Following Bottou, Curtis, Nocedal

▷ STOCHASTIC METHODS

$\min_x f(x)$, $\quad f(x) = \mathbb{E}\, F(x;\xi)$ $\quad$ SA "Stochastic Approx." Robbins Monroe, Polyak

$f(x) = \frac{1}{N}\sum f_i(x)$, $f_i = F(x;\xi_{[i]})$ ← SAA, "Sample Avg Approx" or "ERM"

SGD

$x_{k+1} = x_k - t_k d_k$, $\quad \mathbb{E}(d_k) = \nabla f(x_k)$

(ex: $d_k = \nabla_i f(x_k)$ for $i \sim$ uniform $\{1,\ldots,N\}$)

Assume

- ∇f is L-Lipschitz
- f is μ-Polyak-Łojasiewicz (eg., μ-strongly cvx)
- f bdd below (eg. $f \geq 0$)
- $\mathbb{E}[\|d_k\|^2] \leq M + \overset{=1}{M_G}\|\nabla f(x_k)\|^2$, $\overset{=1}{M_G \geq 1}$ (ie., $M_G = 1$ is possible assumptn)

Thm 1, fixed stepsize (Thm 4.6 Bottou)

let $t_k = t \leq \frac{1}{L\cdot M_G}$, then — or $\frac{1}{L}$ if $M_G = 1$

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{tLM}{2\mu} + (1 - t\mu)^{k-1}\left(f(x_1) - f^* - \frac{tLM}{2\mu}\right).$$

ie., converge quickly to ~~[illegible]~~ region of sol'n.

proof is similar to PL proof, since just use bounds...

Thm 2, diminishing stepsizes (Thm 4.7 Bottou) — ie., $\beta = \frac{2}{\mu}$ is a good choice

let $t_k = \frac{\beta}{\gamma + k}$ for $\beta > \frac{1}{\mu}$, $\gamma > 0$, and $t_1 \leq \frac{1}{L\cdot M_G}$ (or $\frac{1}{L}$ if $M_G = 1$)

Then $\mathbb{E}[f(x_k) - f^*] \leq \frac{\nu}{\gamma + k}$, $\nu = \nu(\beta,\gamma)$ is a constant

|— Minibatching —|

Exploits GPU, and CPU — draw memory hierarchy.

See MATLAB demo

* $\boxed{\text{if } f \text{ is } \mu\text{-strongly cvx, } f(x_k) - f^* \leq \varepsilon \Rightarrow \tfrac{1}{2}\|x - x^*\|^2 \leq \mu^{-1}\varepsilon}$

$\Rightarrow$ So $\mathbb{E}(\|x_k - x^*\|^2) \leq O\left(\frac{1}{k}\right)$

▷ VARIANCE REDUCTION ("Gradient Aggregation")

specific to $f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$, e.g., $f_i(x) = \varphi(a_i^T x - b_i)$ for GLM

<u>Algo: SAGA</u>

```
Initialize: $X^{(i)} = X_0 \quad \forall i = 1, \dots, N$   (each $x^{(i)}$ is n-dimensional)
    and store $\{\nabla f_i(x^{(i)})\}_{i=1}^{N}$ in a $n \times N$ table *
For $k = 1, 2, \dots$
    $j \sim \text{Uniform}([1, \dots, N])$
    $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x^{(i)})$   via table
    $x_{k+1} = x_k - t_k(\nabla f_j(x_k) - \nabla f_j(x^{(j)}) + \bar{z})$
    $x^{(j)} \leftarrow x_k$, update table w/ $\nabla_j f(x^{(j)})$. (update $\bar{z}$)
```

* (see left page: if then can store just $a_i^T x^{(i)} - b_i$

<u>Thm</u> for appropriate t, this converges linearly!

<u>Algo: SVRG</u>

```
For $k = 1, 2, \dots$   "epoch"
    $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_k)$   Full pass through data
    $w_0 \leftarrow x_k$
    For $\ell = 1, 2, \dots, m-1$
        $j \sim \text{Uniform}([1, \dots, N])$
        $w_{\ell+1} = w_\ell - t(\nabla f_j(\cancel{x} w_\ell) - \nabla f_j(x_k) + \bar{z})$

    Option I     $x_{k+1} = w_m$
    Option II    $x_{k+1} = w_\ell$ for $\ell \sim \text{Uni}(\{0, \dots, T-1\})$
    Option III   $x_{k+1} = \frac{1}{m}\sum_{\ell=0}^{m} w_{\ell+1}$
```

<u>Thm</u> For appropriate t, this converges linearly!

▷ ITERATE AVERAGING

First approach, iterate SGD as usual, $x_{k+1} = x_k - t_k \cdot d_k$

but hope $\bar{x}_k := \frac{1}{k}\sum_{j=1}^{k} x_j$ converges faster than x_j.

If we use $t_k = O(1/k)$, it doesn't help.

but, if strongly cvx, choose $t_k = O(\frac{1}{k^\alpha})$ for $\alpha \in (\frac{1}{2}, 1)$

then $\mathbb{E}(\|\bar{x}_k - x^*\|^2) = O(\frac{1}{k})$ while $\mathbb{E}(\|x_k - x^*\|^2) = O(\frac{1}{k^\alpha})$

↗ but with right α, this has better constants. "optimal"

Helps if ill-conditioned.

see "Robust SA" Nemirovsky

"Primal-Dual Avg" Nesterov §6

▷ STEP-SIZES (B.B., Wolfe, Armijo...)

$c_1 \simeq 10^{-4}$

lfe Cond.

- *Sufficient Decrease* (Armijo) $f(x_k + t_k \cdot d_k) \le f(x_k) + c_1 \cdot t_k \cdot \langle d_k, p_k \rangle$

- Prevent short-steps ($t_k \to 0$) by either

  directional deriv.

  a) Curvature Conditions: $\langle \nabla f(x_k + t_k d_k), d_k \rangle \geq c_2 \cdot \langle \nabla f(x_k), d_k \rangle$

  ↙ 2 $c_2 \in (c_1, 1)$ (0.1 to 0.9)

  or Strong Wolfe: Armijo and $|\langle \ldots \rangle| \le c_2 |\langle \ldots, \ldots \rangle|$

  b) backtrack

Goldstein is another possibility, not for quasi-Newton methods