Natasha 2 (Allen-zhu)

Charlie Hou

Problem

- Online stochastic nonconvex optimization

$$\min_{x\in\mathbb{R}^d}\left\{f(x)\stackrel{\text{def}}{=}\mathbb{E}_i[f_i(x)]=\frac{1}{n}\sum_{i=1}^n f_i(x)\right\}$$

- Need to avoid saddle points
  - Random Perturbation
  - Is there another strategy?

Use the Hessian (kind of)!

- When at saddle, take negative curvature direction

Figure 1: Local minimum (left), saddle point (right) and its negative-curvature direction.

- Use Oja's Algorithm to find this direction, which empirically takes roughly twice the time of a gradient computation

Algorithm specified

- Very similar to SVRG and repeatSVRG

Algorithm 2 an informal version of $\texttt{Natasha2}(f, y_0, \varepsilon, \delta)$

Input: function $f(x)$ satisfying Problem (5.1), starting vector y_0, target accuracy $\varepsilon > 0$ and $\delta > 0$.

for $k \leftarrow 0$ **to** ∞ **do**
    Apply Oja's algorithm to find minEV v of $\nabla^2 f(y_k)$.
    if $v \in \mathbb{R}^d$ is found s.t. $v^\top \nabla^2 f(y_k) v \le -\frac{\delta}{2}$ **then**
        $y_{k+1} \leftarrow y_k \pm \frac{\delta}{L_2} v$ where the sign is random.
    else ⋄ *it satisfies* $\nabla^2 f(y_k) \succeq -\delta \mathbf{I}$
        $F(x) = F^k(x) \stackrel{\text{def}}{=} f(x) + L(\max\{0, \|x - y_k\| - \frac{\delta}{L_2}\})^2$.
        $y_{k+1} \leftarrow \texttt{Natasha1.5}(F, y_k, \varepsilon^{-2}, 1, \varepsilon^{4/3}/\delta^{1/3})$
        Break the for loop if have performed $\Theta(\frac{\delta^{1/3}}{\varepsilon^{4/3}})$ first-order steps.
    end if
end for
return y_k.

Guarantees

- What's the weird stuff? Additive terms, reusing previous iterates in the additive terms
  - Natasha 1.5: to bound approximation error, to make it nice
  - Natasha 2: to ensure that grad(F) ~ 0 => grad(f) ~ 0, to make it nice

Theorem 2 (informal). *Under (A1), (A2) and (A4),* Natasha2 *outputs a point* x^{out} *with*

$$\|\nabla f(x^{\mathsf{out}})\| \leq \varepsilon \quad \textit{and} \quad \nabla^2 f(x^{\mathsf{out}}) \succeq -\delta \mathbf{I}$$

in gradient complexity[12]

$$T = \widetilde{O}\left(\frac{1}{\delta^5} + \frac{1}{\delta\varepsilon^3} + \frac{1}{\varepsilon^{3.25}}\right) ,$$

if we hide L, L_2, Δ_f, *and* $\mathcal{V}$ *in the big-O notion.*