

# 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

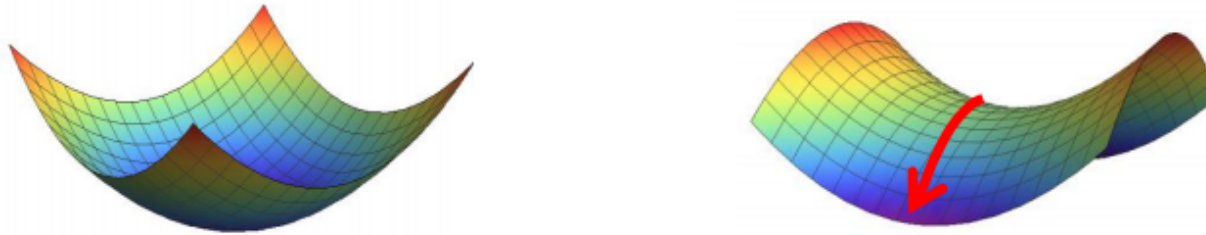


Figure 1: Local minimum (left), saddle point (right) and its negative-curvature direction.

- 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 Natasha2( $f, y_0, \varepsilon, \delta$ )

---

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

```
1: for  $k \leftarrow 0$  to  $\infty$  do
2:   Apply Oja's algorithm to find minEV  $v$  of  $\nabla^2 f(y_k)$ .
3:   if  $v \in \mathbb{R}^d$  is found s.t.  $v^\top \nabla^2 f(y_k) v \leq -\frac{\delta}{2}$  then
4:      $y_{k+1} \leftarrow y_k \pm \frac{\delta}{L_2} v$  where the sign is random.
5:   else  $\diamond$  it satisfies  $\nabla^2 f(y_k) \succeq -\delta \mathbf{I}$ 
6:      $F(x) = F^k(x) \stackrel{\text{def}}{=} f(x) + L(\max\{0, \|x - y_k\| - \frac{\delta}{L_2}\})^2$ .
7:      $y_{k+1} \leftarrow \text{Natasha1.5}(F, y_k, \varepsilon^{-2}, 1, \varepsilon^{4/3}/\delta^{1/3})$ 
8:     Break the for loop if have performed  $\Theta(\frac{\delta^{1/3}}{\varepsilon^{4/3}})$  first-order steps.
9:   end if
10: end for
11: 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  $\text{grad}(F) \sim 0 \Rightarrow \text{grad}(f) \sim 0$ , to make it nice

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

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

*in gradient complexity<sup>12</sup>*

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

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