Surrogate Losses for Online Learning of Stepsizes in Stochastic Non-Convex Optimization

Zhenxun Zhuang[1], Ashok Cutkosky[2], Francesco Orabona[1,3]

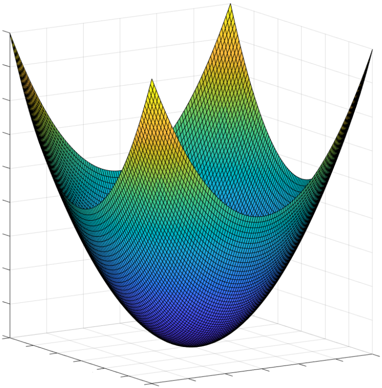
[1]Department of Computer Science, Boston University

[2]Google

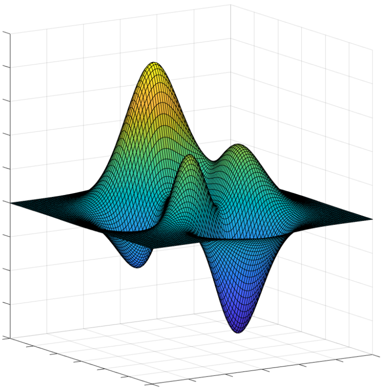
[3]Department of Electrical & Computer Engineering, Boston University

Convex vs. Non-Convex Functions

A Convex Function



A Non-Convex Function



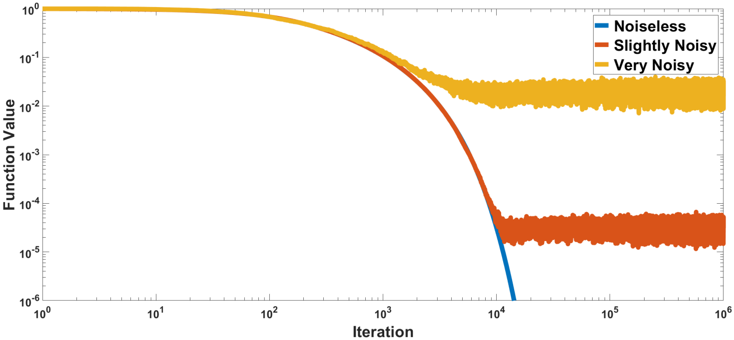
Stationary points: $\|\nabla f(\boldsymbol{x})\| = 0$

Gradient Descent vs. Stochastic Gradient Descent

Gradient Descent: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \nabla f(\boldsymbol{x}_t)$

SGD: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t) \quad \text{with} \quad \mathbb{E}_t[\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)] = \nabla f(\boldsymbol{x}_t)$

Curse of Constant Stepsize



- Ghadimi & Lan (2013): running SGD on M-smooth functions with $\eta \leq \frac{1}{M}$ and assuming $\mathbb{E}_t\left[\|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t) - \nabla f(\boldsymbol{x}_t)\|^2\right] \leq \sigma^2$ yields
$$\mathbb{E}[\|\nabla f(\boldsymbol{x}_i)\|^2] \leq O\left(\frac{f(\boldsymbol{x}_1) - f^\star}{\eta T} + \eta\sigma^2\right) .$$
- Ward et al. (2018) and Li & Orabona (2019) eliminated the need to know $f^\star$ and σ for getting optimal rate by AdaGrad global stepsizes.

Transform Non-Convexity to Convexity by Surrogate Losses

When the objective function is M-smooth, drawing two independent stochastic gradients in each round of SGD, we have (*assume for now* η_t *only depends on past gradients*) :

$$\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)\right] &\leq \mathbb{E}\left[\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\rangle + \frac{M}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2\right] \\
&= \mathbb{E}\left[\langle \nabla f(\boldsymbol{x}_t), -\eta_t \boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\rangle + \frac{M}{2}\eta_t^2\|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2\right] \\
&= \mathbb{E}\left[-\eta_t\langle \boldsymbol{g}(\boldsymbol{x}_t, \xi_t), \boldsymbol{g}(\boldsymbol{x}_t, \xi_t')\rangle + \frac{M\eta_t^2}{2}\|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2\right] .
\end{aligned}$$

Transform Non-Convexity to Convexity by Surrogate Losses

We define the **surrogate loss** for f at round t as

$$\ell_t(\eta) \triangleq -\eta\langle \boldsymbol{g}(\boldsymbol{x}_t, \xi_t), \boldsymbol{g}(\boldsymbol{x}_t, \xi'_t)\rangle + \frac{M\eta^2}{2}\|\boldsymbol{g}(\boldsymbol{x}_t, \xi_t)\|^2 \ .$$

The inequality of last page becomes

$$\mathbb{E}\left[f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)\right] \leq \mathbb{E}\left[\ell_t(\eta_t)\right],$$

which, after summing from $t = 1$ to T gives us:

$$f^\star - f(\boldsymbol{x}_1) \leq \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\ell_t(\eta_t) - \ell_t(\eta)\right]}_{\text{Regret of } \eta_t \text{ wrt optimal } \eta} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\ell_t(\eta)\right]}_{\text{Cumulative loss of optimal } \eta} .$$

SGD with Online Learning

Algorithm 1 Stochastic Gradient Descent with Online Learning (SGDOL)

Input: $\boldsymbol{x}_1 \in \mathcal{X}$, M, an online learning algorithm $\mathcal{A}$

for $t = 1, 2, \ldots, T$ **do**

    Compute η_t by running $\mathcal{A}$ on

    $\ell_i(\eta) = -\eta\langle \boldsymbol{g}(\boldsymbol{x}_i, \xi_i), \boldsymbol{g}(\boldsymbol{x}_i, \xi_i')\rangle + \frac{M\eta^2}{2}\|\boldsymbol{g}(\boldsymbol{x}_i, \xi_i)\|^2, \quad i = 1, \ldots, t-1$

    Receive two independent unbiased estimates of $\nabla f(\boldsymbol{x}_t)$: $\boldsymbol{g}(\boldsymbol{x}_t, \xi_t), \boldsymbol{g}(\boldsymbol{x}_t, \xi_t')$

    Update $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{g}_t$

end for

Output: uniformly randomly choose a $\boldsymbol{x}_k$ from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$.

BOSTON UNIVERSITY

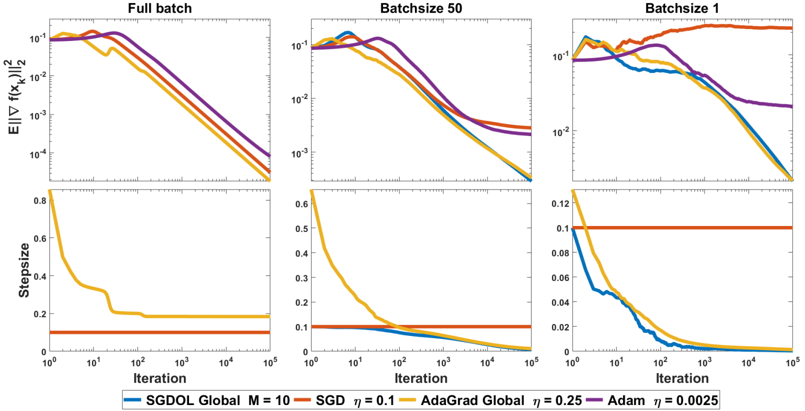
Main Theorem

Theorem 1: Assume some conditions, and make some choice of the online learning algorithm in Algorithm 1, for a smooth function and an uniformly randomly picked $\boldsymbol{x}_k$ from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$, we have:

$$\mathbb{E}\left[\|\nabla f(\boldsymbol{x}_k)\|^2\right] \leq \tilde{\mathcal{O}}\left(\frac{1}{T} + \frac{\sigma}{\sqrt{T}}\right),$$

where $\tilde{\mathcal{O}}$ hides some logarithmic factors.

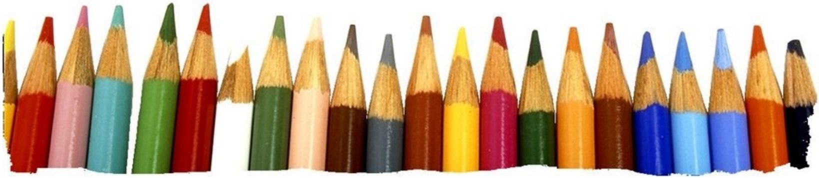
Classification Problem



Objective Function: $\frac{1}{m}\sum_{i=1}^{m}\phi(\boldsymbol{a}_i^\top \boldsymbol{x} - y_i)$ with $\phi(\theta) = \frac{\theta^2}{1+\theta^2}$ on the adult (a9a) training dataset.

THANK YOU!

For more information,
see our poster tonight
@ Pacific Ballroom #105



BOSTON UNIVERSITY