Implicit Gradient Transport

NeurIPS 2019 in Vancouver, Canada

Problem

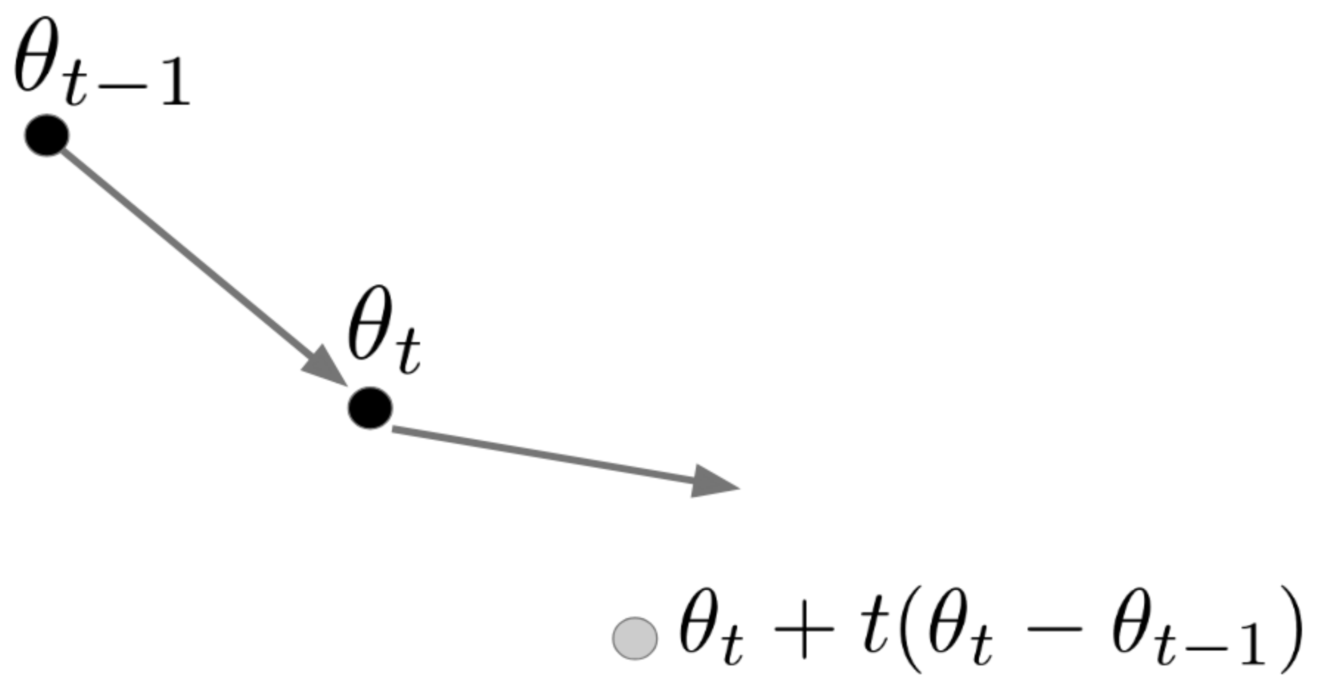
- We’re interested in online stochastic optimization.
- Gradient and accelerated methods do not converge due to stochastic gradients.
- SAG & co. are convergent, but not suited for the online setting.
- **Can we design a *simple* method that converges for this setting ?**

$$\theta_{t+1} = \theta_t - \eta g_t$$

$$g_t = \nabla_{\theta_t} \mathcal{L}(\theta_t)$$

Method

- **Yes!**
- **Big Idea** Transport the *gradient information* from one parameter iterate to another.
- **Concretely** Compute gradient at a shifted point, and average it with previous gradient estimate.
- You get a variance-reduced stochastic gradient, readily **pluggable into any gradient method**.
(e.g. Heavyball, Adam)



$$\gamma_t = \frac{t}{t+1}$$

$$g_t = \gamma_t g_{t-1} + (1-\gamma_t)\hat{g}_t$$

$$\hat{g}_t = \nabla \mathscr{L}(\theta_t + t(\theta_t - \theta_{t-1}))$$

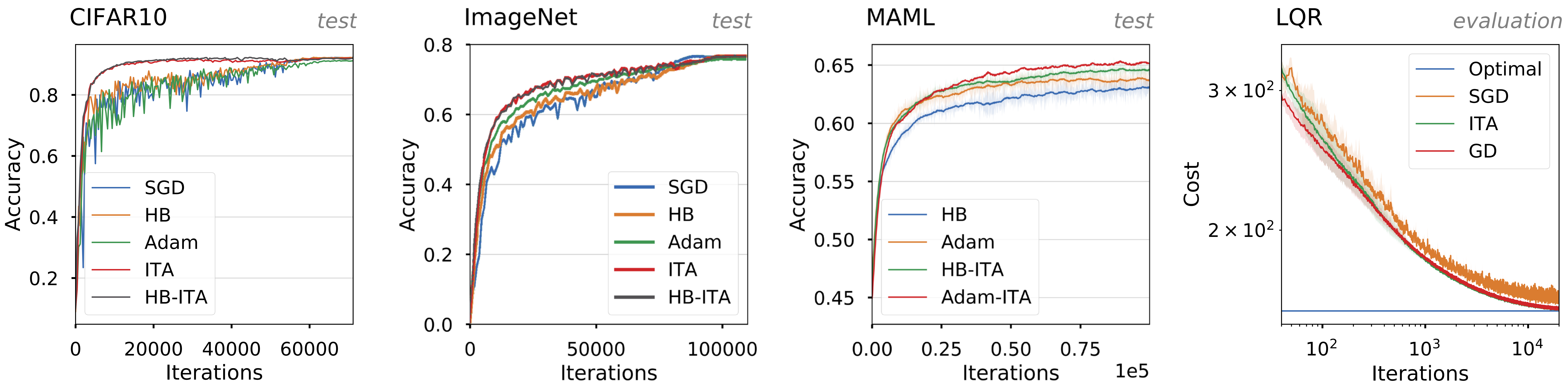
Theory

- **Theorem 1** Plugged into SGD, the IGT gradient estimator converges at a rate of $\mathcal{O}(1/t)$.

- **Theorem 2** Plugged into Heavyball, the IGT gradient estimator achieves the accelerated rate $\mathcal{O}(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^t)$.

- **Caveat** Those results are only proved for quadratic $\mathcal{L}(\theta_t)$.

Experiments



Thank You

Reducing the variance in online optimization by transporting past gradients.

Learn more at bit.ly/31ySnEC or talk to us at Poster #2887, Tuesday 5:30pm.

Séb Arnold

Pierre-Antoine Manzagol

Reza Babanezhad

Ioannis Mitliagkas

Nicolas Le Roux