AISTATS 2018, Playa Blanca, Lanzarote, Spain

Zeroth-order Optimization in High Dimensions

Yining Wang
Carnegie Mellon University

Joint work with Simon Du, Sivaraman Balakrishnan and Aarti Singh

BACKGROUND

- Optimization: $\min_{x \in \mathcal{X}} f(x)$
- Classical setting (first-order):
  - f is known (e.g., a likelihood function or an NN objective)
  - $\nabla f(x)$ can be evaluated, or unbiasedly approximated
- Zeroth-order setting:
  - f is unknown, or very complicated
  - $\nabla f(x)$ is unknown, or very difficult to evaluate.

APPLICATIONS

- Hyper-parameter tuning
  - f maps hyper-parameter x to system performance $f(x)$.
- Experimental design
  - f maps experimental setting to experimental results.
- Communication-efficient optimization
  - Data defining the objective scattered throughout machines
  - Communicating $\nabla f(x)$ is expensive, but $f(x)$ ok.

Formulation

- Convexity: the objective f is convex.
- Noisy observation model:

$$y_t = f(x_t) + \xi_t, \qquad \xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

- Evaluation measure:
  - Simple regret: $f(\widehat{x}_{T+1}) - f^*$
  - Cumulative regret: $\sum_{t=1}^{T} f(x_t) - f^*$

METHODS

- Classical method: ***Estimating Gradient Descent*** (EGD)
- Gradient descent / Mirror descent:

$$x_{t+1} \leftarrow x_t - \eta_t \widehat{g}_t(x_t)$$

$$x_{t+1} \in \arg\min_{z \in \mathbb{R}^d} \{\eta_t \langle \widehat{g}_t(x_t), z\rangle + \Delta_\psi(z, x_t)\}$$

- Estimating gradient:
  - $\widehat{g}_t(x_t) = \frac{d}{\delta} \cdot \mathbb{E}[f(x_t + \delta v_t) v_t]$
  - Gained popularity from (Nemirovski & Yudin'83, Flaxman et al.'05)

Methods

- Classical method: ***Estimating Gradient Descent*** (EGD)
- Gradient descent / Mirror descent: $\widehat{g}_t(x_t) \approx \nabla f(x_t)$

$$x_{t+1} \leftarrow x_t - \eta_t \widehat{g}_t(x_t)$$

$$x_{t+1} \in \arg\min_{z \in \mathbb{R}^d} \{\eta_t \langle \widehat{g}_t(x_t), z \rangle + \Delta_\psi(z, x_t)\}$$

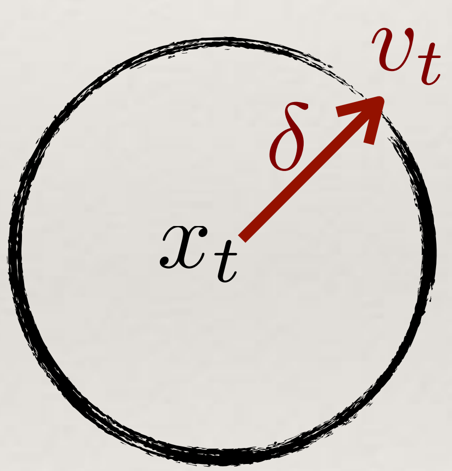
- Estimating gradient:
  - $\widehat{g}_t(x_t) = \frac{d}{\delta} \cdot \mathbb{E}[f(x_t + \delta v_t) v_t]$
  - Gained popularity from (Nemirovski & Yudin'83, Flaxman et al.'05)

Methods

- Classical method: ***Estimating Gradient Descent*** (EGD)
- Gradient descent / Mirror descent: $\widehat{g}_t(x_t) \approx \nabla f(x_t)$

  $$x_{t+1} \leftarrow x_t - \eta_t \widehat{g}_t(x_t)$$

  $$x_{t+1} \in \arg\min_{z \in \mathbb{R}^d} \{\eta_t \langle \widehat{g}_t(x_t), z \rangle + \Delta_\psi(z, x_t)\}$$



- Estimating gradient:
  - $\widehat{g}_t(x_t) = \dfrac{d}{\delta} \cdot \mathbb{E}[f(x_t + \delta v_t) v_t]$
  - Gained popularity from (Nemirovski & Yudin'83, Flaxman et al.'05)

Methods

- Classical method: ***Estimating Gradient Descent*** (EGD)
- Classical analysis
  - Supposing $\|\nabla f\| \leq H$ and $\|x^*\|_* \leq B$
  - Stochastic GD/MD: $f(\widehat{x}) - f^* \lesssim BH/\sqrt{T}$
  - Estimating GD/MD: $f(\widehat{x}) - f^* \lesssim \sqrt{d \cdot BH}/T^{1/4}$
- Problem: cannot exploit (sparse) structure in x^*

Methods

- Classical method: ***Estimating Gradient Descent*** (EGD)
- Classical analysis $\mathbb{E}\widehat{g}_t(x_t) = \nabla f(x_t)$
  - Supposing $\|\nabla f\| \leq H$ and $\|x^*\|_* \leq B$

First-order

  - Stochastic GD/MD: $f(\widehat{x}) - f^* \lesssim BH/\sqrt{T}$
  - Estimating GD/MD: $f(\widehat{x}) - f^* \lesssim \sqrt{d \cdot BH}/T^{1/4}$
- Problem: cannot exploit (sparse) structure in x^*

METHODS

- Classical method: ***Estimating Gradient Descent*** (EGD)
- Classical analysis $\mathbb{E}\widehat{g}_t(x_t) = \nabla f(x_t)$
  - Supposing $\|\nabla f\| \leq H$ and $\|x^*\|_* \leq B$

First-order

  - Stochastic GD/MD: $f(\widehat{x}) - f^* \lesssim BH/\sqrt{T}$
  - Estimating GD/MD: $f(\widehat{x}) - f^* \lesssim \sqrt{d \cdot BH}/T^{1/4}$

Zeroth-order

- Problem: cannot exploit (sparse) structure in x^*

$\mathbb{E}\|\widehat{g}_t(x_t) - \nabla f(x_t)\|_2^2$ small, but

$\mathbb{E}\widehat{g}_t(x_t) \neq \nabla f(x_t)$

Methods

- Classical method: ***Estimating Gradient Descent*** (EGD)
- Classical analysis $\mathbb{E}\widehat{g}_t(x_t) = \nabla f(x_t)$
  - Supposing $\|\nabla f\| \leq H$ and $\|x^*\|_* \leq B$

First-order

  - Stochastic GD/MD: $f(\widehat{x}) - f^* \lesssim BH/\sqrt{T}$
  - Estimating GD/MD: $f(\widehat{x}) - f^* \lesssim \sqrt{d \cdot BH}/T^{1/4}$

Zeroth-order

- Problem: cannot exploit (sparse) structure in x^*

$\mathbb{E}\|\widehat{g}_t(x_t) - \nabla f(x_t)\|_2^2$ small, but

$\mathbb{E}\widehat{g}_t(x_t) \neq \nabla f(x_t)$

ASSUMPTIONS

- The "function sparsity" assumption:

$$f(x) \equiv f(x_S) \qquad\qquad S \subseteq [d], |S| = s \ll d$$

- Strong theoretically, but slightly acceptable in practice
  - Hyper-parameter tuning: performance not sensitive to many input parameters
  - Visual stimuli optimization: most brain activities are not related to visual reactions.

Lasso gradient estimate

- Local linear approximation:

$$f(x_t + \delta v_t) \approx f(x_t) + \delta \langle \nabla f(x_t), v_t \rangle$$

- Lasso gradient estimate:
  - Sample $v_1, \cdots, v_n$ and observe $y_i \approx f(x_t + \delta v_i) - f(x_t)$
  - Construct a **sparse linear system:**

$$\widetilde{Y} = Y/\delta = V \nabla f(x_t) + \varepsilon$$

Lasso gradient estimate

- Local linear approximation:

$$f(x_t + \delta v_t) \approx f(x_t) + \delta \langle \nabla f(x_t), v_t \rangle$$

- Lasso gradient estimate:
  - Construct a **sparse linear system:**

$$\widetilde{Y} = Y/\delta = V \nabla f(x_t) + \varepsilon$$

  - Because $\nabla f(x_t)$ is **sparse**, one can use the **Lasso**

$$\widehat{g}_t(x_t) \in \arg\min_{g \in \mathbb{R}^d} \left\{ \|\widetilde{Y} - Vg\|_2^2 + \lambda \|g\|_1 \right\}$$

Lasso gradient estimate

- Local linear approximation:

$$f(x_t + \delta v_t) \approx f(x_t) + \delta \langle \nabla f(x_t), v_t \rangle$$

- Lasso gradient estimate:
  - Construct a **sparse linear system:**

$$\widetilde{Y} = Y/\delta = V \nabla f(x_t) + \varepsilon$$

certain "de-biasing" required … see paper

  - Because $\nabla f(x_t)$ is **sparse**, one can use the **Lasso**

$$\widehat{g}_t(x_t) \in \arg\min_{g \in \mathbb{R}^d} \left\{ \|\widetilde{Y} - Vg\|_2^2 + \lambda \|g\|_1 \right\}$$

Main results

Theorem. Suppose $f(x) \equiv f(x_S)$ for some $|S| = s \ll d$, and other smoothness conditions on f hold. Then

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f^* \lesssim \text{poly}(s, \log d) \cdot T^{-1/4}$$

Furthermore, for smoother f the $T^{-1/4}$ can be improved to $T^{-1/3}$

Main results

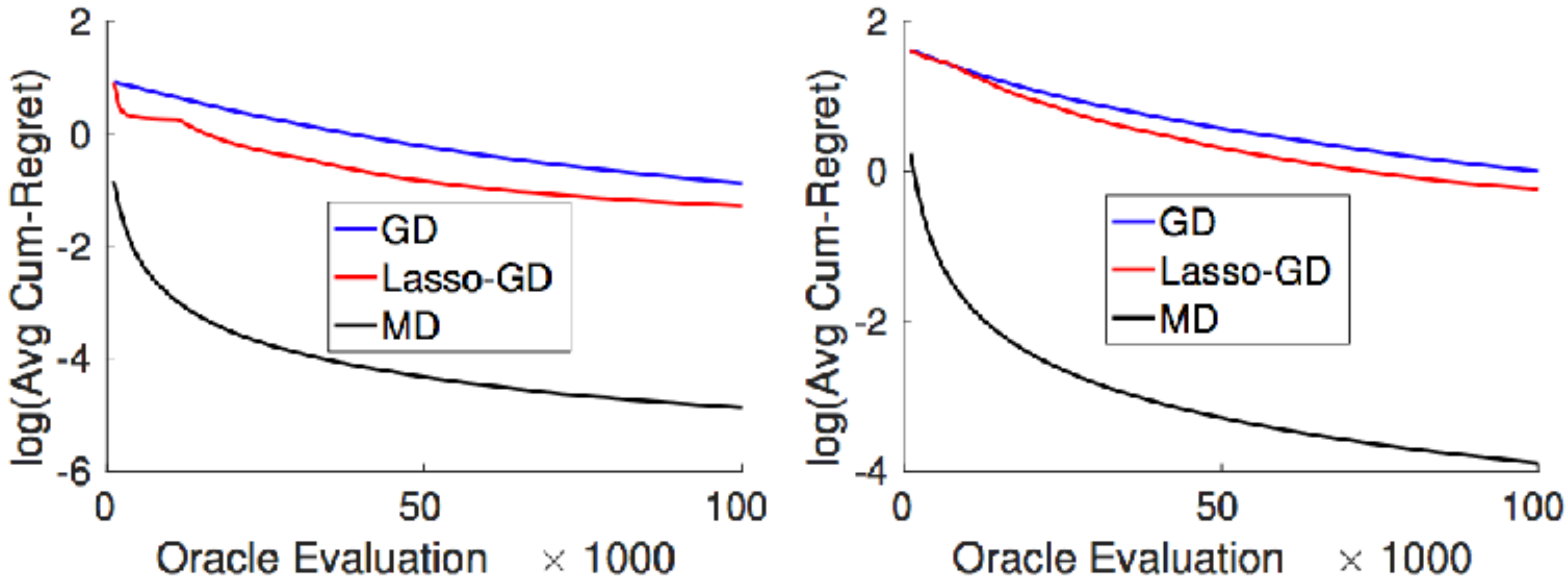
Theorem. Suppose $f(x) \equiv f(x_S)$ for some $|S| = s \ll d$, and other smoothness conditions on f hold. Then

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f^* \lesssim \mathrm{poly}(s, \log d) \cdot T^{-1/4}$$

Furthermore, for smoother f the $T^{-1/4}$ can be improved to $T^{-1/3}$

can handle "high-dimensional" setting $d \gg T$

Simulation results



(a) $s = 10, d = 100$ (b) $s = 20, d = 100$

Figure 1: Sparse quadratic optimization with identity quadratic term.

SIMULATION RESULTS

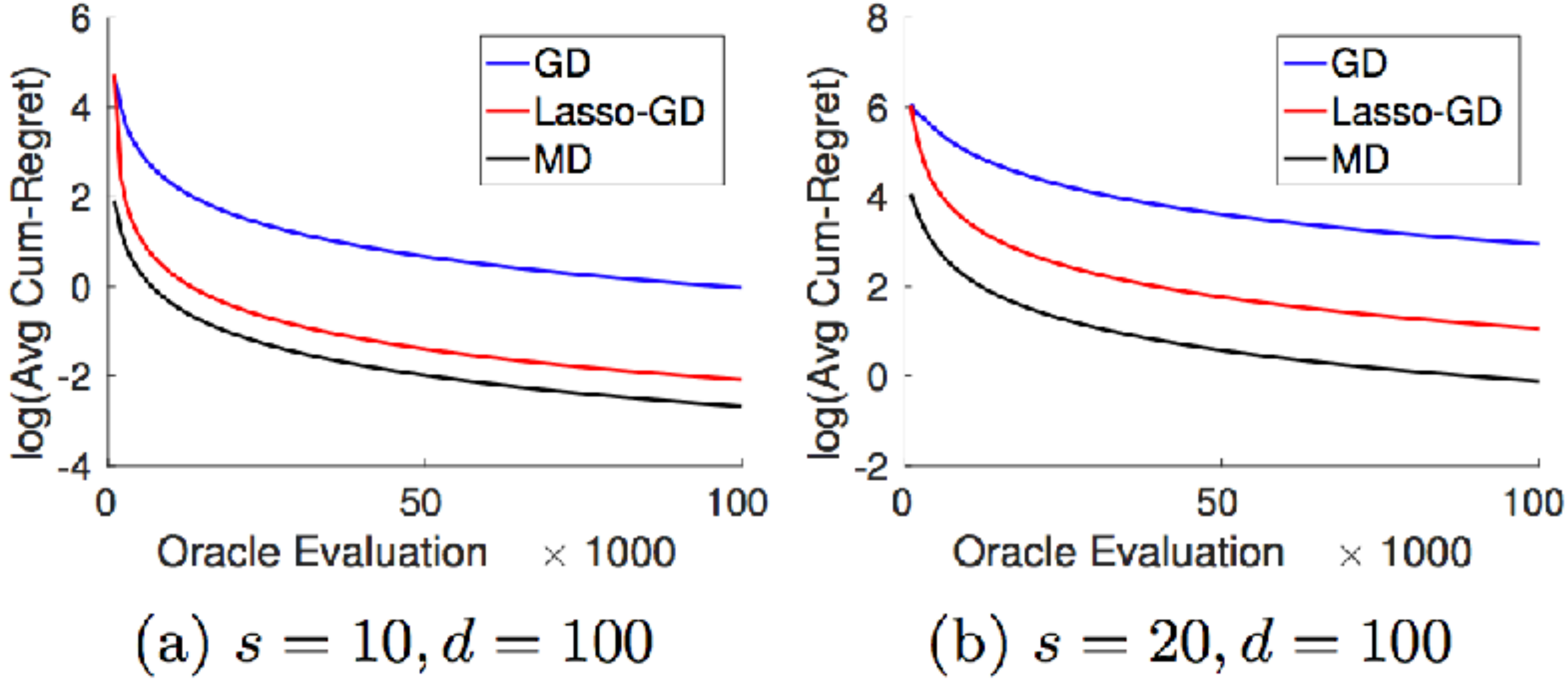


Figure 3: Sparse fourth-degree polynomial optimization with identity quadratic term.

Open questions

- Is function/gradient sparsity absolutely necessary?
  - Recall in **first-order** case, only **solution** x^* sparsity required
  - More specifically, only need $\|x^*\|_1 \le B, \|\nabla f\|_\infty \le H$
  - **Conjecture**: if f only satisfies the above condition, then

$$\inf_{\widehat{x}_T} \sup_f \mathbb{E}\left[f(\widehat{x}_T) - f^*\right] \gtrsim \mathrm{poly}(d, 1/T)$$

OPEN QUESTIONS

- Is $T^{-1/2}$ convergence achievable in high dimensions?
  - Challenge 1: MD is awkward in exploiting strong convexity:

$$f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle + \frac{\nu^2}{2} \Delta_\psi(x', x)$$

  - Challenge 2: the Lasso gradient estimate is less efficient — can we design convex body K such that

$$\widehat{g}_t(x_t) = \frac{\rho(K)}{\delta} \int_{\partial K} f(x_t + \delta v) \mathbf{n}(v) \mathrm{d}\mu(v)$$

is a good gradient estimator in high dimensions?

OPEN QUESTIONS

Wish to replace with
$\|x' - x\|_1^2$

- Is $T^{-1/2}$ convergence achievable in high dimensions?
  - Challenge 1: MD is awkward in exploiting strong convexity:

$$f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle + \frac{\nu^2}{2} \boxed{\Delta_\psi(x', x)}$$

  - Challenge 2: the Lasso gradient estimate is less efficient — can we design convex body K such that

$$\widehat{g}_t(x_t) = \frac{\rho(K)}{\delta} \int_{\partial K} f(x_t + \delta v)\mathbf{n}(v) \mathrm{d}\mu(v)$$

is a good gradient estimator in high dimensions?

Thank you!

Questions