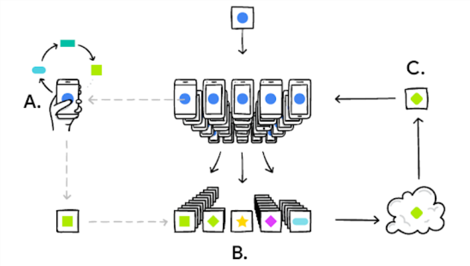


Source - Google

Information-Constrained Optimization:
Can Adaptive Processing of Gradients help?

NeurIPS 2021

Jayadev Acharya, Cornell University
Clément Canonne, University of Sydney
Prathamesh Mayekar, Indian Institute of Science
Himanshu Tyagi, Indian Institute of Science

1. The Setup

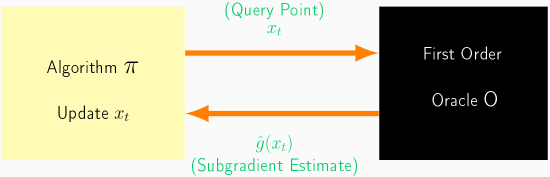
Classical Setup:[1] The query complexity framework

Algorithm π

Algorithm π: Minimize unknown function f using an oracle O.

[1] Nemirovsky, A. S., and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.

Classical Setup:[1] The query complexity framework

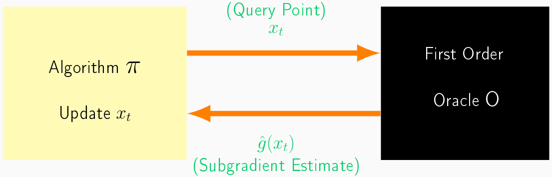


Algorithm π: Minimize unknown function f using an oracle O.

Oracle O: Output noisy sub-gradient estimate $\hat{g}(x_t)$ for query x_t.

[1] Nemirovsky, A. S., and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.

Classical Setup:[1] The query complexity framework



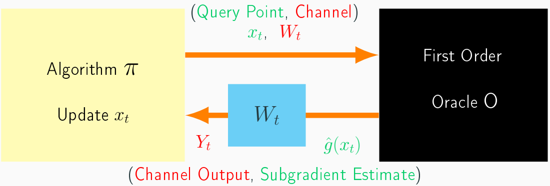
Algorithm π: Minimize unknown function f using an oracle O.

Oracle O: Output noisy sub-gradient estimate $\hat{g}(x_t)$ for query x_t.

What is the best possible convergence rate?

[1] Nemirovsky, A. S., and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.

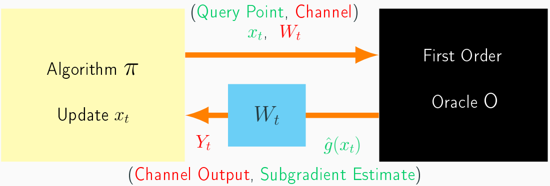
Our Setup: Optimization under information constraints



$\hat{g}(x_t)$ can be sent using a channel $W_t \in W$ and only the output Y_t available to the algorithm.

$$Y_t \mid \hat{g}(x_t) \sim W(\cdot \mid \hat{g}(x_t)), \text{ where } W_t \in \mathcal{W}.$$

Our Setup: Optimization under information constraints

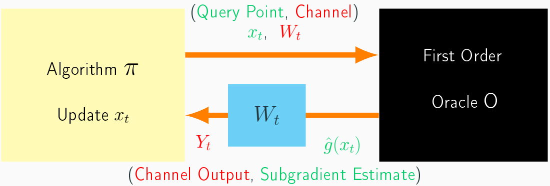


$\hat{g}(x_t)$ can be sent using a channel $W_t \in W$ and only the output Y_t available to the algorithm.

$$Y_t \mid \hat{g}(x_t) \sim W(\cdot \mid \hat{g}(x_t)), \text{ where } W_t \in \mathcal{W}.$$

Reduces to classical setup if we are allowed the identity channel.

Our Setup: Optimization under information constraints



$\hat{g}(x_t)$ can be sent using a channel $W_t \in W$ and only the output Y_t available to the algorithm.

$$Y_t \mid \hat{g}(x_t) \sim W(\cdot \mid \hat{g}(x_t)), \text{ where } W_t \in \mathcal{W}.$$

Reduces to classical setup if we are allowed the identity channel.

Information-constraints

1. Local Differential Privacy: The family $\mathcal{W}_{\mathtt{priv},\varepsilon}$ comprising W s.t.

$$\ln \frac{W(y \mid x)}{W(y \mid x')} \leq \varepsilon \quad \forall x, x' \in \mathcal{X}, y \in \mathcal{Y}.$$

Information-constraints

1. Local Differential Privacy: The family $\mathcal{W}_{\mathtt{priv},\varepsilon}$ comprising W s.t.

$$\ln \frac{W(y \mid x)}{W(y \mid x')} \leq \varepsilon \quad \forall x, x' \in \mathcal{X}, y \in \mathcal{Y}.$$

2. Communication Constraints: the family $\mathcal{W}_{\mathtt{com},r}$ comprising W s.t.

(the output range) $|\mathcal{Y}| \leq 2^r$.

Information-constraints

1. Local Differential Privacy: The family $\mathcal{W}_{\texttt{priv},\varepsilon}$ comprising W s.t.

$$\ln \frac{W(y \mid x)}{W(y \mid x')} \le \varepsilon \quad \forall x, x' \in \mathcal{X}, y \in \mathcal{Y}.$$

2. Communication Constraints: the family $\mathcal{W}_{\texttt{com},r}$ comprising W s.t.

(the output range) $|\mathcal{Y}| \le 2^r$.

3. Random Coordinate Descent: the family $\mathcal{W}_{\texttt{obl}}$ comprising W s.t. for any input $g \in \mathbb{R}^d$,

W outputs $\{g(I), I\}$ where I is a randomly chosen coordinate.

Function Family

- Domain $\mathcal{X}$ will be a set of diameter D in $\mathbb{R}^d$.

Function Family

- Domain $\mathcal{X}$ will be a set of diameter D in $\mathbb{R}^d$.
- Function, Oracle class $\mathcal{O}$ consists of all tuples $\{f, O\}$ such that
  1. f is *convex*.

Function Family

- Domain $\mathcal{X}$ will be a set of diameter D in $\mathbb{R}^d$.
- Function, Oracle class $\mathcal{O}$ consists of all tuples $\{f, O\}$ such that
  1. f is *convex*.
  2. *Unbiased*: $\mathbb{E}\left[\hat{g}(x)|x\right] \in \partial f(x)$.

Function Family

- Domain $\mathcal{X}$ will be a set of diameter D in $\mathbb{R}^d$.
- Function, Oracle class $\mathcal{O}$ consists of all tuples $\{f, O\}$ such that
  1. f is *convex*.
  2. *Unbiased*: $\mathbb{E}\left[\hat{g}(x)|x\right] \in \partial f(x)$.
  3. *Almost surely norm-bounded*: $\|\hat{g}(x)\|_2 \leq B$.

Function Family

- Domain $\mathcal{X}$ will be a set of diameter D in $\mathbb{R}^d$.
- Function, Oracle class $\mathcal{O}$ consists of all tuples $\{f, O\}$ such that
  1. f is *convex*.
  2. *Unbiased*: $\mathbb{E}\left[\hat{g}(x)|x\right] \in \partial f(x)$.
  3. *Almost surely norm-bounded*: $\|\hat{g}(x)\|_2 \leq B$.

Our Goal:

- Characterize

$$\mathcal{E}(T, \mathcal{W}) := \inf_{\pi \in \Pi_T} \inf_{\{W_t\}_{t \in [T]} \in \mathcal{W}} \sup_{\{f, O\} \in \mathcal{O}} \mathbb{E}\left[f(x(\pi, Q))\right] - f^*$$

Worst-case gap to optimality using "joint-best" T query optimization algo and coding scheme.

Function Family

- Domain $\mathcal{X}$ will be a set of diameter D in $\mathbb{R}^d$.
- Function, Oracle class $\mathcal{O}$ consists of all tuples $\{f, O\}$ such that
  1. f is *convex*.
  2. *Unbiased*: $\mathbb{E}[\hat{g}(x)|x] \in \partial f(x)$.
  3. *Almost surely norm-bounded*: $\|\hat{g}(x)\|_2 \leq B$.

Our Goal:

- Characterize

$$\mathcal{E}(T, \mathcal{W}) := \inf_{\pi \in \Pi_T} \inf_{\{W_t\}_{t \in [T]} \in \mathcal{W}} \sup_{\{f, O\} \in \mathcal{O}} \mathbb{E}[f(x(\pi, Q))] - f^*$$

  Worst-case gap to optimality using "joint-best" T query optimization algo and coding scheme.

- Classical Result (no channel constraints): $\mathcal{E}(T) = \Theta\left(\frac{DB}{\sqrt{T}}\right)$.

2. Lower Bounds for Information-Constrained Optimization

Lower Bounds

For T large enough,

(Private Optimization) and $\varepsilon \in [0, 1]$,

$$\mathcal{E}(T, \mathcal{W}_{\mathtt{priv},\varepsilon}) \geq \frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{\varepsilon^2}}.$$

Lower Bounds

For T large enough,

(Private Optimization) and $\varepsilon \in [0, 1]$,

$$\mathcal{E}(T, \mathcal{W}_{\texttt{priv},\varepsilon}) \geq \frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{\varepsilon^2}}.$$

(Communication-Constrained Optimization) and $r \in [d]$,

$$\mathcal{E}(T, \mathcal{W}_{\texttt{com},r}) \geq \frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{r}}.$$

Lower Bounds

For T large enough,

(Private Optimization) and $\varepsilon \in [0,1]$,

$$\mathcal{E}(T, \mathcal{W}_{\mathtt{priv},\varepsilon}) \geq \frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{\varepsilon^2}}.$$

(Communication-Constrained Optimization) and $r \in [d]$,

$$\mathcal{E}(T, \mathcal{W}_{\mathtt{com},r}) \geq \frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{r}}.$$

(Random Coordinate descent)

$$\mathcal{E}(T, \mathcal{W}_{\mathtt{obl}}) \geq \frac{DB}{\sqrt{T}} \cdot \sqrt{d}.$$

Lower Bounds

For T large enough,

(Private Optimization) and $\varepsilon \in [0, 1]$,

$$\mathcal{E}(T, \mathcal{W}_{\mathtt{priv},\varepsilon}) \geq \frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{\varepsilon^2}}.$$

(Communication-Constrained Optimization) and $r \in [d]$,

$$\mathcal{E}(T, \mathcal{W}_{\mathtt{com},r}) \geq \frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{r}}.$$

(Random Coordinate descent)

$$\mathcal{E}(T, \mathcal{W}_{\mathtt{obl}}) \geq \frac{DB}{\sqrt{T}} \cdot \sqrt{d}.$$

These LBs are tight and achieved by nonadaptive gradient coding.

3. Proof

Proof: The difficult case for convex family

Similar to [Nemirovski, Yudin 83], [Agarwal, Bartlett, Ravikumar, Wainwright 12].

Proof: The difficult case for convex family

1. Domain: $\mathcal{X} = \frac{D}{2\sqrt{d}}[-1,1]^d$.
2. Difficult subclass of functions and oracle:

$$f_V(x) := \frac{B\delta}{\sqrt{d}} \sum_{i=1}^{d} V(i)x(i),\ \hat{g}_t(i) = \begin{cases} +B/\sqrt{d} & \text{w.p. } (1+\delta V(i))/2 \\ -B/\sqrt{d} & \text{w.p. } (1-\delta V(i))/2 \end{cases}$$

where $V \sim \mathrm{Uniform}\{-1,+1\}^d$.

3. Average Mutual Information Bound:

$$\boxed{\mathbb{E}\left[f_V(x_T)\right] - f^* \geq \frac{BD\delta}{4}\left(1 - \sqrt{\frac{2}{d}\sum_{i=1}^{d} I(V(i) \wedge \{Y_t\}_{t\in[T]})}\right).}$$

Proof: The difficult case for convex family

1. Domain: $\mathcal{X} = \frac{D}{2\sqrt{d}}[-1,1]^d$.
2. Difficult subclass of functions and oracle:

$$f_V(x) := \frac{B\delta}{\sqrt{d}} \sum_{i=1}^{d} V(i)x(i),\ \hat{g}_t(i) = \begin{cases} +B/\sqrt{d} & \text{w.p. } (1+\delta V(i))/2 \\ -B/\sqrt{d} & \text{w.p. } (1-\delta V(i))/2 \end{cases}$$

where $V \sim \mathrm{Uniform}\{-1,+1\}^d$.

3. Average Mutual Information Bound:

$$\boxed{\mathbb{E}\left[f_V(x_T)\right] - f^* \geq \frac{BD\delta}{4}\left(1 - \sqrt{\frac{2}{d}\sum_{i=1}^{d} I(V(i) \wedge \{Y_t\}_{t\in[T]})}\right).}$$

Proof: The difficult case for convex family

1. Domain: $\mathcal{X} = \frac{D}{2\sqrt{d}}[-1,1]^d$.
2. Difficult subclass of functions and oracle:

$$f_V(x) := \frac{B\delta}{\sqrt{d}} \sum_{i=1}^{d} V(i)x(i),\ \hat{g}_t(i) = \begin{cases} +B/\sqrt{d} & \text{w.p. } (1+\delta V(i))/2 \\ -B/\sqrt{d} & \text{w.p. } (1-\delta V(i))/2 \end{cases}$$

where $V \sim \mathrm{Uniform}\{-1,+1\}^d$.

3. Average Mutual Information Bound:

$$\boxed{\mathbb{E}\left[f_V(x_T)\right] - f^* \geq \frac{BD\delta}{4}\left(1 - \sqrt{\frac{2}{d}\sum_{i=1}^{d} I(V(i) \wedge \{Y_t\}_{t\in[T]})}\right).}$$

Proof: The difficult case for convex family

1. Domain: $\mathcal{X} = \frac{D}{2\sqrt{d}}[-1,1]^d$.
2. Difficult subclass of functions and oracle:

$$f_V(x) := \frac{B\delta}{\sqrt{d}} \sum_{i=1}^{d} V(i)x(i),\ \hat{g}_t(i) = \begin{cases} +B/\sqrt{d} & \text{w.p. } (1+\delta V(i))/2 \\ -B/\sqrt{d} & \text{w.p. } (1-\delta V(i))/2 \end{cases}$$

where $V \sim \mathrm{Uniform}\{-1,+1\}^d$.

3. Average Mutual Information Bound:

$$\boxed{\mathbb{E}\left[f_V(x_T)\right] - f^* \geq \frac{BD\delta}{4}\left(1 - \sqrt{\frac{2}{d}\sum_{i=1}^{d} I(V(i) \wedge \{Y_t\}_{t\in[T]})}\right).}$$

Prior work gets a larger mutual information: $I(V \wedge \{Y_t\}_{t\in[T]})$.

Proof: The difficult case for convex family

1. Domain: $\mathcal{X} = \frac{D}{2\sqrt{d}}[-1,1]^d$.
2. Difficult subclass of functions and oracle:

$$f_V(x) := \frac{B\delta}{\sqrt{d}} \sum_{i=1}^{d} V(i)x(i),\ \hat{g}_t(i) = \begin{cases} +B/\sqrt{d} & \text{w.p. } (1+\delta V(i))/2 \\ -B/\sqrt{d} & \text{w.p. } (1-\delta V(i))/2 \end{cases}$$

where $V \sim \mathrm{Uniform}\{-1,+1\}^d$.

3. Average Mutual Information Bound:

$$\boxed{\mathbb{E}\left[f_V(x_T)\right] - f^* \geq \frac{BD\delta}{4}\left(1 - \sqrt{\frac{2}{d}\sum_{i=1}^{d} I(V(i) \wedge \{Y_t\}_{t\in[T]})}\right).}$$

Prior work gets a larger mutual information: $I(V \wedge \{Y_t\}_{t\in[T]})$.

Can bound average mutual information using: [J Acharya, C Canonne, Z Sun, and H Tyagi, "Unified lower bounds for interactive high-dimensional estimation under information constraints," 2020]

Lower bounds for strongly convex functions are trickier

1. Quadratic functions of the form $\|x - \mu_v\|_2^2$ are the bottlenecks.

Lower bounds for strongly convex functions are trickier

1. Quadratic functions of the form $\|x - \mu_v\|_2^2$ are the bottlenecks.
2. The gradients are no longer independent of the queried point – introduces complicated correlation between V and the coded sequence of gradients.

Lower bounds for strongly convex functions are trickier

1. Quadratic functions of the form $\|x - \mu_v\|_2^2$ are the bottlenecks.
2. The gradients are no longer independent of the queried point – introduces complicated correlation between V and the coded sequence of gradients.
3. Need upper bounds on MI for adaptive protocols to even prove nonadaptive lower bounds.

Lower bounds for strongly convex functions are trickier

1. Quadratic functions of the form $\|x - \mu_v\|_2^2$ are the bottlenecks.
2. The gradients are no longer independent of the queried point – introduces complicated correlation between V and the coded sequence of gradients.
3. Need upper bounds on MI for adaptive protocols to even prove nonadaptive lower bounds.
4. Even here, adaptive gradient coding does not help.

Lower bounds for strongly convex functions are trickier

1. Quadratic functions of the form $\|x - \mu_v\|_2^2$ are the bottlenecks.
2. The gradients are no longer independent of the queried point – introduces complicated correlation between V and the coded sequence of gradients.
3. Need upper bounds on MI for adaptive protocols to even prove nonadaptive lower bounds.
4. Even here, adaptive gradient coding does not help.

So, is there a class of optimization problem where adaptivity helps?

4. Adaptivity Helps!

A structured optimization problem

For $\mathcal{X} = [-1,1]^d$, consider

$$\min_{\mathcal{X}} \|x - v\|_2^2 \,,$$

where $v \in [-1,1]^d$ is s-block sparse. That is,

1. Only one of the block $\{is, \ldots, i(s+1)\}$ of coordinates can have non-zero values.
2. All the coordinates in a block have the same absolute value.

Oracle: Outputs $2(x_t - Z_t)$,
where $\{Z_t\}_{t=1}^{\infty}$ is i.i.d., $Z_1 = \{-1,1\}^d$, and $\mathbb{E}\left[Z_1\right] = v$.

Channel Constraint: Algorithm can only see one coordinate of the gradient estimate. (RCD channel family)

The gap between adaptive and nonadaptive protocols

Lower bound for nonadaptive protocols:

For any nonadaptive protocol we can find a block-sparse v s.t.

$$\mathbb{E}\left[\|x_T - v\|_2^2\right] \geq \frac{ds}{T}.$$

Upper bound for a adaptive protocol:

There exist an adaptive protocol such that

$$\mathbb{E}\left[\|x_T - v\|_2^2\right] \lessapprox \frac{d + s^2}{T}.$$

The Adaptive Protocol

1. **(Exploration Phase):** Use the first $T/2$ queries to find the non-sparse block.
    1.1 Sample a representative coordinate from each block $Ts/2d$ times.
    1.2 Select the block with absolute largest sample mean.
2. **(Exploitation Phase):** Use the last $T/2$ iterations to sample all the s coordinates within the chosen block $T/2s$ times.
3. **(Final Estimate):**
    3.1 For all the coordinates outside the chosen block set the mean estimate to be 0.
    3.2 For all the coordinates within the chosen block use the sample mean estimate.

In Summary ...

1. Adaptive gradient coding doesn't help for standard optimization.

In Summary ...

1. Adaptive gradient coding doesn't help for standard optimization.

2. For structured optimization problems, adaptive coding does help.

In Summary ...

1. Adaptive gradient coding doesn't help for standard optimization.
2. For structured optimization problems, adaptive coding does help.
3. Adaptive techniques useful for proving nonadaptive lower bounds.

In Summary . . .

1. Adaptive gradient coding doesn't help for standard optimization.
2. For structured optimization problems, adaptive coding does help.
3. Adaptive techniques useful for proving nonadaptive lower bounds.

Thank You!