RATQ: A Universal Fixed-Length Quantizer for Stochastic Optimization

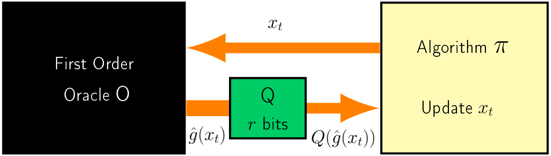
Prathamesh Mayekar

Joint work with

Himanshu Tyagi

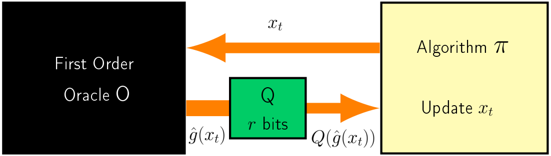
Department of ECE,
Indian Institute of Science

The Setup



$\hat{g}(x_t)$ can be sent to a finite precision r using a Q of our choice.

The Setup



$\hat{g}(x_t)$ can be sent to a finite precision r using a Q of our choice.

Main Question:

Which $\{\pi, Q\}$ gives the best convergence rate for r bits and T queries?

Assumptions

- Domain $\mathcal{X}$ will be the Euclidean ball of diameter D in $\mathbb{R}^d$.

Assumptions

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

Assumptions

- Domain $\mathcal{X}$ will be the Euclidean ball 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*: $\mathbb{E}\left[\hat{g}(x)|x\right] \in \partial f(x)$.

Assumptions

- Domain $\mathcal{X}$ will be the Euclidean ball 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*: $\|\hat{g}(x)\|_2 \leq B$.

Assumptions

- Domain $\mathcal{X}$ will be the Euclidean ball 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

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

worst-case gap to optimality using " joint-best"
T query optimzation algo and r bit quantizer.

Assumptions

- Domain $\mathcal{X}$ will be the Euclidean ball 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

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

  worst-case gap to optimality using " joint-best" T query optimzation algo and r bit quantizer.

- Classical Result: $\mathcal{E}(T, \infty) = \Theta\left(\frac{DB}{\sqrt{T}}\right)$

Lower bound

Theorem

$$\mathcal{E}(T,r) \geq \Omega\left(\frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge r}}\right)$$

Lower bound

Theorem

$$\mathcal{E}(T, r) \geq \Omega\left(\frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge r}}\right)$$

Goal: Matching upper bound for any precision constraint r.

Lower bound

Theorem

$$\mathcal{E}(T, r) \geq \Omega\left(\frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge r}}\right)$$

Goal: Matching upper bound for any precision constraint r.

Previous work: PSGD +

Lower bound

Theorem

$$\mathcal{E}(T,r) \geq \Omega\left(\frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge r}}\right)$$

Goal: Matching upper bound for any precision constraint r.

Previous work: PSGD +

- Uniform Quantization $\Rightarrow$ upper bound off by $\sqrt{\log d}$.

Lower bound

Theorem

$$\mathcal{E}(T, r) \geq \Omega\left(\frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge r}}\right)$$

Goal: Matching upper bound for any precision constraint r.

Previous work: PSGD +

- Uniform Quantization $\Rightarrow$ upper bound off by $\sqrt{\log d}$.
- Random rotation based quantizer in [Suresh et al. 17] $\Rightarrow$ upper bound off by $\sqrt{\log \log d}$.

Lower bound

Theorem

$$\mathcal{E}(T, r) \geq \Omega\left(\frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge r}}\right)$$

Goal: Matching upper bound for any precision constraint r.

Previous work: PSGD +

- Uniform Quantization $\Rightarrow$ upper bound off by $\sqrt{\log d}$.
- Random rotation based quantizer in [Suresh et al. 17] $\Rightarrow$ upper bound off by $\sqrt{\log \log d}$.
- Variable length quantizers in [Alistarh et al. 17] and [Suresh et al. 17] $\Rightarrow$ upper 'can' be off by $\sqrt{\log d}$.

Lower bound

Theorem

$$\mathcal{E}(T, r) \geq \Omega\left(\frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge r}}\right)$$

Goal: Matching upper bound for any precision constraint r.

Previous work: PSGD +

- Uniform Quantization $\Rightarrow$ upper bound off by $\sqrt{\log d}$.
- Random rotation based quantizer in [Suresh et al. 17] $\Rightarrow$ upper bound off by $\sqrt{\log \log d}$.
- Variable length quantizers in [Alistarh et al. 17] and [Suresh et al. 17] $\Rightarrow$ upper 'can' be off by $\sqrt{\log d}$.

We will show a tighter upper bound of $\sqrt{\log \ln^* d}$.

Convergence with compressed gradients

We use Projected Subgradient descent (PSGD) with compressed gradients.

Theorem

Consider an unbiased, r-bit quantizer Q. *Then,*

$$\mathcal{E}(T, r) \leq \frac{D \cdot \alpha(Q)}{\sqrt{T}},$$

where $\alpha(Q) := \sup_{Y \in \mathbb{R}^d : \|Y\|_2^2 \leq B^2 \text{ a.s.}} \sqrt{\underbrace{\mathbb{E}\left[\|Q(Y) - Y\|_2^2\right]}_{MSE} + B^2}.$

Convergence with compressed gradients

We use Projected Subgradient descent (PSGD) with compressed gradients.

Theorem

Consider an unbiased, r-bit quantizer Q. Then,

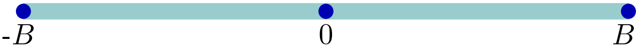
$$\mathcal{E}(T, r) \leq \frac{D \cdot \alpha(Q)}{\sqrt{T}},$$

where $\alpha(Q) := \sup_{Y \in \mathbb{R}^d : \|Y\|_2^2 \leq B^2 \text{ a.s.}} \sqrt{\underbrace{\mathbb{E}\left[\|Q(Y) - Y\|_2^2\right]}_{MSE} + B^2}.$

Find the minimum MSE Quantizer of the ℓ_2 ball
s.t. 1) Precision is r bits, 2) Unbiased.

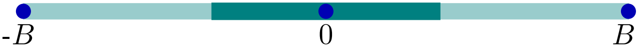
RATQ: Our quantizer for the ℓ_2 ball

Input to RATQ: Y such that $\|Y\|_2 \leq B$.



RATQ: Our quantizer for the ℓ_2 ball

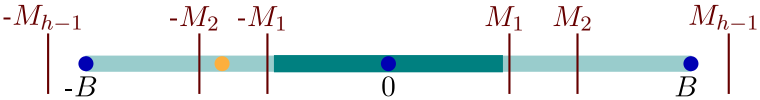
Input to RATQ: Y such that $\|Y\|_2 \leq B$.



1. Rotate Y using randomized Hadamard transform.
   - Leads to each coordinate being subgaussian with a variance factor $\frac{B^2}{d}$, instead of B^2.

RATQ: Our quantizer for the ℓ_2 ball

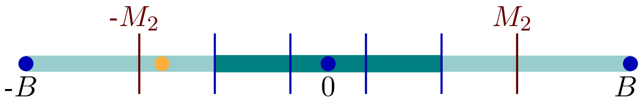
Input to RATQ: Y such that $\|Y\|_2 \leq B$.



1. Rotate Y using randomized Hadamard transform.
   - Leads to each coordinate being subgaussian with a variance factor $\frac{B^2}{d}$, instead of B^2.
2. For each coordinate, choose smallest one of the intervals $[-M_i, M_i]$ containing it and quantize uniformly to k levels.

RATQ: Our quantizer for the ℓ_2 ball

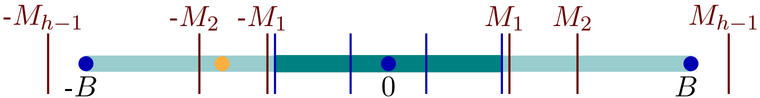
Input to RATQ: Y such that $\|Y\|_2 \le B$.



1. Rotate Y using randomized Hadamard transform.
   - Leads to each coordinate being subgaussian with a variance factor $\frac{B^2}{d}$, instead of B^2.
2. For each coordinate, choose smallest one of the intervals $[-M_i, M_i]$ containing it and quantize uniformly to k levels.

RATQ: Our quantizer for the ℓ_2 ball

Input to RATQ: Y such that $\|Y\|_2 \leq B$.



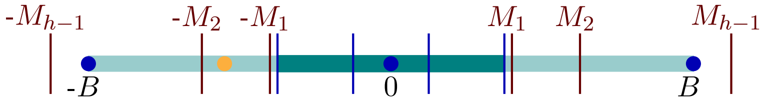
1. Rotate Y using randomized Hadamard transform.
   - Leads to each coordinate being subgaussian with a variance factor $\frac{B^2}{d}$, instead of B^2.
2. For each coordinate, choose smallest one of the intervals $[-M_i, M_i]$ containing it and quantize uniformly to k levels.
3. Per coordinate precision is $\log h + \log k$ bits.

   Per coordinate MSE $\approx \frac{1}{(k-1)^2} \sum_{i \in [h]} M_i^2 \cdot p(M_{i-1})$,

   $p(M)$ is the prob. of the absolute value exceeding M.

RATQ: Our quantizer for the ℓ_2 ball

Input to RATQ: Y such that $\|Y\|_2 \leq B$.



1. Rotate Y using randomized Hadamard transform.
   - Leads to each coordinate being subgaussian with a variance factor $\frac{B^2}{d}$, instead of B^2.
2. For each coordinate, choose smallest one of the intervals $[-M_i, M_i]$ containing it and quantize uniformly to k levels.
3. Per coordinate precision is $\log h + \log k$ bits.

   Per coordinate MSE $\approx \frac{1}{(k-1)^2} \sum_{i \in [h]} M_i^2 \cdot p(M_{i-1})$,

   $p(M)$ is the prob. of the absolute value exceeding M.
4. $M_{i+1}^2 \approx e^{M_i^2} (tetration)$.

Complete characterization of $\mathcal{E}(T, r)$

- RATQ uses a per coordinate precision of $\log \ln^* d$ to get $\alpha(Q) = O(B)$.

Complete characterization of $\mathcal{E}(T, r)$

- RATQ uses a per coordinate precision of $\log \ln^* d$ to get $\alpha(Q) = O(B)$.
- To make it work for r $(< d)$ bits, uniformly sample $\frac{r}{\log \ln^* d}$ coordinates; $\alpha(Q) = O\left(B \cdot \sqrt{\frac{d \log \ln^* d}{r}}\right)$.

Theorem

$$\frac{c_0 DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge r}} \leq \mathcal{E}(T, r) \leq \frac{c_1 DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge \frac{r}{\log \ln^* d}}}$$

Mean Square Bounded Oracles

ℓ_2 Case: Instead of $\|\hat{g}(x)\|_2^2 \leq B^2$, we have $\mathbb{E}\left[\|\hat{g}(x)\|_2^2\right] \leq B^2$.

Mean Square Bounded Oracles

ℓ_2 Case: Instead of $\|\hat{g}(x)\|_2^2 \leq B^2$, we have $\mathbb{E}\left[\|\hat{g}(x)\|_2^2\right] \leq B^2$.

We may not be able to find an unbiased quantizer.

Mean Square Bounded Oracles

ℓ_2 Case: Instead of $\|\hat{g}(x)\|_2^2 \leq B^2$, we have $\mathbb{E}\left[\|\hat{g}(x)\|_2^2\right] \leq B^2$.

We may not be able to find an unbiased quantizer.

Use a *gain-shape* quantizer:

- express $y \equiv (\|y\|_2, y/\|y\|_2)$ and quantize each part separately

Mean Square Bounded Oracles

ℓ_2 Case: Instead of $\|\hat{g}(x)\|_2^2 \leq B^2$, we have $\mathbb{E}\left[\|\hat{g}(x)\|_2^2\right] \leq B^2$.

We may not be able to find an unbiased quantizer.

Use a *gain-shape* quantizer:

- express $y \equiv (\|y\|_2, y/\|y\|_2)$ and quantize each part separately

"The gain quantizer must be carefully chosen"

Mean Square Bounded Oracles

ℓ_2 Case: Instead of $\|\hat{g}(x)\|_2^2 \leq B^2$, we have $\mathbb{E}\left[\|\hat{g}(x)\|_2^2\right] \leq B^2$.

We may not be able to find an unbiased quantizer.

Use a *gain-shape* quantizer:

- express $y \equiv (\|y\|_2, y/\|y\|_2)$ and quantize each part separately

"The gain quantizer must be carefully chosen"

Specifically, uniform quantizers have the following bottleneck:

- To attain $DB/\sqrt{T}$, r must exceed $d + \log T$
- We construct a "heavy-tailed" oracle for these bounds

Mean Square Bounded Oracles

ℓ_2 Case: Instead of $\|\hat{g}(x)\|_2^2 \leq B^2$, we have $\mathbb{E}\left[\|\hat{g}(x)\|_2^2\right] \leq B^2$.

We may not be able to find an unbiased quantizer.

Use a *gain-shape* quantizer:

- express $y \equiv (\|y\|_2, y/\|y\|_2)$ and quantize each part separately

"The gain quantizer must be carefully chosen"

Specifically, uniform quantizers have the following bottleneck:

- To attain $DB/\sqrt{T}$, r must exceed $d + \log T$
- We construct a "heavy-tailed" oracle for these bounds

RATQ combined with "adaptive geometric" gain quantizer requires $\approx d + \log \log T$ bits to attain $DB/\sqrt{T}$ rate.

Concluding Remarks

Our quantizers:

- RATQ with PSGD attains optimal convergence for fixed precision upto a $\sqrt{\log \ln^* d}$ factor
- A gain-shape variant of RATQ comes close to the optimal for mean square bounded oracles.

Our lower bounds:

- For mean square bounded oracles:
  - lower bound by constructing heavy tailed oracles

Thank You!