# Gradient Sampling Methods with Inexact Subproblem Solves and Gradient Aggregation

**Frank E. Curtis**, Lehigh University

joint work with

**Minhan Li**, Lehigh University

presented at

SIAM Conference on Computational Science and Engineering

March 4, 2021

# Outline

# Outline

# Locally Lipschitz optimization

Consider optimization problems of the form

$$\min_{x \in \mathbb{R}^n} \ f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is

- locally Lipschitz over $\mathbb{R}^n$;
- continuously differentiable on an open set $\mathcal{D}_f$ that has full measure in $\mathbb{R}^n$.

Our goal is to improve upon the **gradient sampling** methodology.

# Main idea

If $f$ is smooth, then the steepest descent direction at $x_k$ is $-\nabla f(x_k)$ since

$$\min_{\|d\|_2 \leq 1} f'(x_k, d) = \min_{\|d\|_2 \leq 1} \nabla f(x_k)^T d = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|_2}.$$

# Main idea

If $f$ is smooth, then the steepest descent direction at $x_k$ is $-\nabla f(x_k)$ since

$$\min_{\|d\|_2 \le 1} f'(x_k, d) = \min_{\|d\|_2 \le 1} \nabla f(x_k)^T d = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|_2}.$$

If $f$ is locally Lipschitz, then ideally one can solve

$$\min_{\|d\|_2 \le 1} f^\circ(x_k, d) = \arg\min_{\|d\|_2 \le 1} \left( \max_{g \in \eth f(x_k)} g^T d \right).$$

# Main idea

If $f$ is smooth, then the steepest descent direction at $x_k$ is $-\nabla f(x_k)$ since

$$\min_{\|d\|_2 \leq 1} f'(x_k, d) = \min_{\|d\|_2 \leq 1} \nabla f(x_k)^T d = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|_2}.$$

If $f$ is locally Lipschitz, then ideally one can solve

$$\min_{\|d\|_2 \leq 1} f^\circ(x_k, d) = \arg\min_{\|d\|_2 \leq 1} \left( \max_{g \in \bar{\partial} f(x_k)} g^T d \right).$$

However, this is intractable, so we approximate:

$$\arg\min_{\|d\|_2 \leq 1} \left( \max_{g \in \bar{\partial}_{\epsilon_k} f(x_k)} g^T d \right) \approx \arg\min_{\|d\|_2 \leq 1} \left( \max_{g \in \mathcal{G}_k} g^T d \right) = -\frac{g_k}{\|g_k\|_2},$$

where $g_k$ is the min-norm element of $\mathcal{G}_k := \{\nabla f(x_k), \nabla f(x_{k,1}), \ldots, \nabla f(x_{k,p})\}$.

## Gradient sampling (Burke, Lewis, and Overton)

At a given iterate $x_k \in \mathbb{R}^n$ and with a sampling radius $\epsilon_k \in \mathbb{R}_{>0}$:

- **sample** $p \geq n+1$ points in $\mathbb{B}(x_k, \epsilon_k)$
- **evaluate** $\mathcal{G}_k := \{\nabla f(x_k), \nabla f(x_{k,1}), \dots, \nabla f(x_{k,p})\}$
- **compute** the minimum norm element of $\text{conv}(\mathcal{G}_k)$, call it $g_k$
- **check** $\|g_k\|_2 = \mathcal{O}(\epsilon_k)$; if so, then set $\epsilon_{k+1} < \epsilon_k$; else $\epsilon_{k+1} \leftarrow \epsilon_k$
- **perform** a backtracking line search to obtain $x_k - \alpha_k g_k$
- **perturb** $x_k - \alpha_k g_k \approx x_{k+1}$ (if necessary) to ensure $x_{k+1} \in \mathcal{D}_f$

# Gradient sampling (Burke, Lewis, and Overton)

At a given iterate $x_k \in \mathbb{R}^n$ and with a sampling radius $\epsilon_k \in \mathbb{R}_{>0}$:

- **sample** $p \geq n+1$ points in $\mathbb{B}(x_k, \epsilon_k)$
- **evaluate** $\mathcal{G}_k := \{\nabla f(x_k), \nabla f(x_{k,1}), \dots, \nabla f(x_{k,p})\}$
- **compute** the minimum norm element of $\text{conv}(\mathcal{G}_k)$, call it $g_k$
- **check** $\|g_k\|_2 = \mathcal{O}(\epsilon_k)$; if so, then set $\epsilon_{k+1} < \epsilon_k$; else $\epsilon_{k+1} \leftarrow \epsilon_k$
- **perform** a backtracking line search to obtain $x_k - \alpha_k g_k$
- **perturb** $x_k - \alpha_k g_k \approx x_{k+1}$ (if necessary) to ensure $x_{k+1} \in \mathcal{D}_f$

With probability one, either:

(i) $\{f(x_k)\} \searrow -\infty$ or

(ii) $\{\epsilon_k\} \searrow 0$ and every limit point of $\{x_k\}$ is stationary for $f$.

## Shortcomings and enhancements

Potential shortcomings of the basic algorithm:

- $p \geq n + 1$ gradient evaluations per iteration
- no (approximate) second-order information
- no exploitation of structure of nonsmoothness

Proposed enhancements:

- adaptive sampling (Curtis and Que)
- variable-metric variants (Curtis and Que)
- manifold sampling (Khan, Larson, Menickelly, Wild, Zhou)

# Shortcoming and our contribution

In all of the algorithms mentioned so far:

- QP subproblems have potentially many constraints, and
- QP subproblems need to be solved *exactly* in each iteration

Our contributions:

- *inexact* subproblem solves
- *gradient aggregation* to limit subproblem sizes

...all while maintaining convergence guarantees of the basic method.

# Outline

Motivation

Inexact Subproblem Solutions

Gradient Aggregation

Conclusion

# QP subproblems

Primal-dual form of the gradient sampling QP subproblems:

$$\left\{ \begin{aligned} \min_{(z,d)\in\mathbb{R}\times\mathbb{R}^n} \ & z + \tfrac{1}{2}\|d\|_{H_k}^2 \\ \text{s.t. } & G_k^T d \leq z\mathbb{1} \end{aligned} \right\} \quad \left\{ \begin{aligned} \max_{y\in\mathbb{R}^{p_k+1}} \ & -\tfrac{1}{2}\|G_k y_k\|_{W_k}^2 \\ \text{s.t. } & \mathbb{1}^T y = 1,\ y \geq 0 \end{aligned} \right\}$$

- $p_k$ is the number of gradients available (in addition to $\nabla f(x_k)$)
- $G_k$ is a matrix with gradients as columns
- $H_k$ is a Hessian approximation
- $W_k = H_k^{-1}$ is an inverse Hessian approximation

# QP subproblems

Primal-dual form of the gradient sampling QP subproblems:

$$\left\{\begin{array}{rl}\min\limits_{(z,d)\in\mathbb{R}\times\mathbb{R}^n} & z+\frac{1}{2}\|d\|_{H_k}^2\\ \text{s.t.} & G_k^T d \leq z\mathbb{1}\end{array}\right\} \quad \left\{\begin{array}{rl}\max\limits_{y\in\mathbb{R}^{p_k+1}} & -\frac{1}{2}\|G_k y_k\|_{W_k}^2\\ \text{s.t.} & \mathbb{1}^T y = 1,\ y\geq 0\end{array}\right\}$$

- $p_k$ is the number of gradients available (in addition to $\nabla f(x_k)$)
- $G_k$ is a matrix with gradients as columns
- $H_k$ is a Hessian approximation
- $W_k = H_k^{-1}$ is an inverse Hessian approximation

Given feasible $y_{k,j}$, a corresponding primal feasible solution:

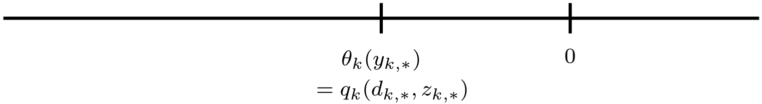
- $d_{k,j} \leftarrow -W_k G_k y_{k,j}$
- $z_{k,j} \leftarrow \max_{i\in\{0,\ldots,p_k\}} \nabla f(x_{k,i})^T d_{k,j}$

As $y_{k,j} \to y_{k,*}$, this converges to primal-dual solution.

## Primal-dual termination test

Consider the primal and dual objective functions:

$$q_k(d,z) = z + \tfrac{1}{2}\|d\|_{H_k}^2 \quad \text{and} \quad \theta_k(y) = -\tfrac{1}{2}\|G_k y\|_{W_k}^2$$

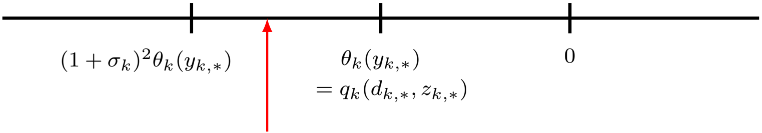


# Primal-dual termination test

Consider the primal and dual objective functions:

$$q_k(d,z) = z + \tfrac{1}{2}\|d\|_{H_k}^2 \quad \text{and} \quad \theta_k(y) = -\tfrac{1}{2}\|G_k y\|_{W_k}^2$$

Given a prescribed inexactness parameter $\sigma_k \in (0,\infty)$:




want to compute $y_{k,j}$

with $\theta_k(y_{k,j})$ in this range

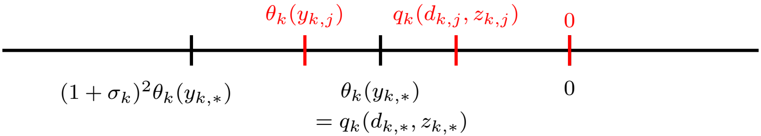
without knowing $\theta_k(y_{k,*})$

# Termination test 1

If the following condition is satisfied

$$\underbrace{q_k(d_{k,j}, z_{k,j}) - \theta_k(y_{k,j})}_{\text{primal-dual gap}} \leq (\sigma_k^2 + 2\sigma_k)\ \underbrace{(0 - q_k(d_{k,j}, z_{k,j}))}_{\text{gap from zero solution}}$$

then the desired condition is satisfied.

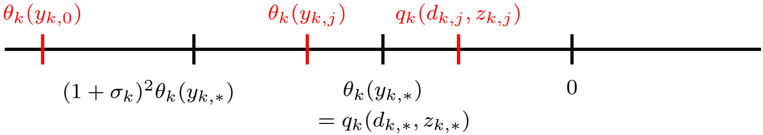


# Termination test 2

If the following condition is satisfied (for some $\rho \in (0,1)$)

$$\underbrace{\theta_k(y_{k,j}) - \theta_k(y_{k,0})}_{\text{dual improvement}} \geq \left( \max\left\{ 1 - \frac{\sigma_k^2 + 2\sigma_k}{\frac{\theta_k(y_{k,0})}{q_k(d_{k,j},z_{k,j})} - 1}, \rho \right\} \right) \underbrace{(q_k(d_{k,j},z_{k,j}) - \theta_k(y_{k,0}))}_{\text{primal improvement}}$$

then the desired condition is satisfied.



## Complete algorithm

The complete algorithm involves

- adaptive sampling (in some iterations, no sampling)
- L-BFGS Hessian approximations
- inexact subproblem solves

# Numerical experiments with NonOpt

Table: Results for GS-exact.

| name | obj | its | f evs | g evs | qp its | CPU |
|---|---|---|---|---|---|---|
| MaxQ | 3.050E-07 | 3717 | 14216 | 5912 | 6859 | 30.55 |
| MxHilb | 1.820E-05 | 526 | 5597 | 4416 | 2006 | 40.79 |
| ChainedLQ | -3.946E+02 | 268 | 4397 | 6631 | 60789 | 35.91 |
| ChainedCB3_1 | 6.180E+02 | 337 | 4858 | 6046 | 45035 | 36.59 |
| ChainedCB3_2 | 4.818E+03 | 95 | 591 | 292 | 598 | 33.88 |
| ActiveFaces | 3.083E-02 | 21 | 669 | 619 | 6173 | 37.15 |
| BrownFunction_2 | 3.347E-03 | 233 | 3647 | 4843 | 31333 | 31.82 |
| ChainedMifflin_2 | -1.550E+02 | 482 | 9991 | 18229 | 160818 | 62.86 |
| ChainedCrescent_1 | 5.197E-03 | 33 | 252 | 201 | 128 | 33.33 |
| ChainedCrescent_2 | 1.258E-03 | 397 | 6441 | 9462 | 77608 | 47.18 |
| Test29_2 | 4.840E-05 | 966 | 9390 | 7096 | 18357 | 38.39 |
| Test29_5 | 9.194E-05 | 508 | 4311 | 2373 | 3736 | 39.07 |
| Test29_6 | 2.263E-04 | 706 | 9305 | 9479 | 40356 | 33.13 |
| Test29_11 | 1.913E+03 | 347 | 5216 | 7693 | 66261 | 35.65 |
| Test29_13 | 1.747E+02 | 338 | 6313 | 10516 | 66438 | 41.17 |
| Test29_17 | 3.961E-05 | 408 | 5341 | 4296 | 14739 | 42.63 |
| Test29_19 | 6.247E-08 | 644 | 7561 | 9105 | 45696 | 43.67 |
| Test29_20 | 1.339E-04 | 1777 | 21947 | 23897 | 110077 | 40.45 |
| Test29_22 | 4.539E-05 | 40 | 574 | 377 | 10453 | 66.28 |
| Test29_24 | 5.562E-05 | 2708 | 49258 | 89275 | 349192 | 76.25 |

# Numerical experiments with NonOpt

Table: Results for GS-inexact.

| name | obj | its | f evs | g evs | qp its | CPU | CPU diff |
|---|---|---|---|---|---|---|---|
| MaxQ | 2.870E-07 | 3863 | 14676 | 6083 | 6121 | 27.75 | -9.18% |
| MxHilb | 2.000E-05 | 464 | 5872 | 4410 | 1835 | 40.50 | -0.69% |
| ChainedLQ | -3.946E+02 | 247 | 3855 | 5510 | 54310 | 34.55 | -3.81% |
| ChainedCB3_1 | 6.180E+02 | 336 | 4712 | 5444 | 31627 | 22.33 | -38.98% |
| ChainedCB3_2 | 4.818E+03 | 94 | 614 | 213 | 387 | 22.94 | -32.28% |
| ActiveFaces | 3.083E-02 | 21 | 669 | 619 | 23 | 1.27 | -96.58% |
| BrownFunction_2 | 3.131E-03 | 256 | 3917 | 4861 | 32461 | 32.34 | 1.65% |
| ChainedMifflin_2 | -1.550E+02 | 442 | 9214 | 16351 | 127904 | 50.16 | -20.21% |
| ChainedCrescent_1 | 4.627E-03 | 34 | 256 | 202 | 123 | 32.01 | -3.96% |
| ChainedCrescent_2 | 1.065E-03 | 332 | 5735 | 8969 | 68458 | 44.08 | -6.58% |
| Test29_2 | 4.961E-05 | 942 | 9384 | 6877 | 16719 | 38.35 | -0.11% |
| Test29_5 | 6.824E-04 | 216 | 1455 | 800 | 1316 | 13.80 | -64.67% |
| Test29_6 | 2.034E-04 | 703 | 9297 | 9493 | 35815 | 29.99 | -9.48% |
| Test29_11 | 1.913E+03 | 433 | 5686 | 7806 | 98615 | 46.45 | 30.28% |
| Test29_13 | 1.747E+02 | 363 | 7356 | 12937 | 71641 | 46.94 | 14.03% |
| Test29_17 | 4.549E-05 | 410 | 5337 | 4237 | 13976 | 41.63 | -2.34% |
| Test29_19 | 3.507E-08 | 580 | 6804 | 8239 | 37114 | 35.94 | -17.71% |
| Test29_20 | 1.158E-04 | 425 | 6222 | 9015 | 42825 | 16.48 | -59.25% |
| Test29_22 | 6.458E-05 | 36 | 526 | 347 | 355 | 2.91 | -95.61% |
| Test29_24 | 4.351E-05 | 2197 | 34581 | 59774 | 287530 | 55.83 | -26.77% |

# Outline

Motivation

Inexact Subproblem Solutions

Gradient Aggregation

Conclusion

# Subgradient aggregation

Subgradient aggregation is a well-known technique for bundle methods.

- It has not previously been used in gradient sampling,
- and generally is harder to employ in nonconvex settings.

However, since it can *drastically* reduce the size of subproblems, it's worth a try.

## Gradient aggregation

Recall the primal-dual form of the gradient sampling QP subproblems:

$$\left\{\begin{array}{c}\min\limits_{(z,d)\in\mathbb{R}\times\mathbb{R}^n} \ z + \frac{1}{2}\|d\|_{H_k}^2 \\ \text{s.t. } G_k^T d \leq z\mathbb{1}\end{array}\right\} \quad \left\{\begin{array}{c}\max\limits_{y\in\mathbb{R}^{p_k+1}} \ -\frac{1}{2}\|G_k y_k\|_{W_k}^2 \\ \text{s.t. } \mathbb{1}^T y = 1,\ y \geq 0\end{array}\right\}$$

At the primal-dual optimal solution:

$$d_{k,*} = -W_k G_k y_{k,*}$$

## Gradient aggregation

Recall the primal-dual form of the gradient sampling QP subproblems:

$$\left\{ \begin{array}{rl} \min\limits_{(z,d)\in\mathbb{R}\times\mathbb{R}^n} & z + \frac{1}{2}\|d\|_{H_k}^2 \\ \text{s.t.} & G_k^T d \le z\mathbb{1} \end{array} \right\} \quad \left\{ \begin{array}{rl} \max\limits_{y\in\mathbb{R}^{p_k+1}} & -\frac{1}{2}\|G_k y_k\|_{W_k}^2 \\ \text{s.t.} & \mathbb{1}^T y = 1,\ y \ge 0 \end{array} \right\}$$

At the primal-dual optimal solution:

$$d_{k,*} = -W_k G_k y_{k,*}$$

Hence, the primal optimal solution is also a solution to

$$\begin{array}{rl} \min\limits_{(z,d)\in\mathbb{R}\times\mathbb{R}^n} & z + \frac{1}{2}\|d\|_{H_k}^2 \\ \text{s.t.} & (G_k y_{k,*})^T d \le z \quad \leftarrow \text{single constraint!} \end{array}$$

## Gradient aggregation

Recall the primal-dual form of the gradient sampling QP subproblems:

$$\left\{ \begin{array}{rl} \min\limits_{(z,d)\in\mathbb{R}\times\mathbb{R}^n} & z + \tfrac{1}{2}\|d\|_{H_k}^2 \\ \text{s.t.} & G_k^T d \leq z\mathbb{1} \end{array} \right\} \quad \left\{ \begin{array}{rl} \max\limits_{y\in\mathbb{R}^{p_k+1}} & -\tfrac{1}{2}\|G_k y_k\|_{W_k}^2 \\ \text{s.t.} & \mathbb{1}^T y = 1,\ y \geq 0 \end{array} \right\}$$

At the primal-dual optimal solution:

$$d_{k,*} = -W_k G_k y_{k,*}$$

Hence, the primal optimal solution is also a solution to

$$\begin{array}{rl} \min\limits_{(z,d)\in\mathbb{R}\times\mathbb{R}^n} & z + \tfrac{1}{2}\|d\|_{H_k}^2 \\ \text{s.t.} & (G_k y_{k,*})^T d \leq z \quad \leftarrow \text{single constraint!} \end{array}$$

If the adaptive sampling strategy is to augment $G_k$, then replace:

$$\underbrace{G_k}_{p_k+1 \text{ columns}} \quad \text{with} \quad \underbrace{G_k y_{k,*}}_{1 \text{ column}}$$

# Numerical experiments with NonOpt

Table: Results for GS-inexact-agg.

| name | obj | its | f evs | g evs | qp its | CPU | CPU diff |
|---|---|---|---|---|---|---|---|
| MaxQ | 2.460E-07 | 3539 | 13171 | 4967 | 5387 | 26.01 | -14.87% |
| MxHilb | 1.115E-04 | 429 | 4184 | 2696 | 1826 | 31.90 | -21.80% |
| ChainedLQ | -3.946E+02 | 229 | 5286 | 6861 | 58896 | 38.35 | 6.79% |
| ChainedCB3_1 | 6.180E+02 | 285 | 5698 | 6630 | 22419 | 16.57 | -54.71% |
| ChainedCB3_2 | 4.818E+03 | 89 | 561 | 238 | 483 | 27.96 | -17.48% |
| ActiveFaces | 3.083E-02 | 21 | 669 | 619 | 23 | 1.22 | -96.72% |
| BrownFunction_2 | 1.843E-03 | 238 | 4533 | 4872 | 19376 | 17.60 | -44.68% |
| ChainedMifflin_2 | -1.550E+02 | 516 | 12762 | 16994 | 187575 | 69.35 | 10.32% |
| ChainedCrescent_1 | 2.795E-03 | 24 | 141 | 66 | 71 | 18.90 | -43.29% |
| ChainedCrescent_2 | 9.704E-04 | 315 | 6123 | 6851 | 29742 | 22.27 | -52.80% |
| Test29_2 | 5.104E-05 | 1108 | 11368 | 6307 | 14573 | 30.91 | -19.49% |
| Test29_5 | 7.822E-05 | 414 | 3768 | 1825 | 3340 | 33.84 | -13.39% |
| Test29_6 | 2.326E-04 | 886 | 12819 | 10279 | 33892 | 28.40 | -14.27% |
| Test29_11 | 1.913E+03 | 324 | 5563 | 5714 | 30196 | 8.62 | -75.82% |
| Test29_13 | 1.747E+02 | 253 | 6419 | 8598 | 34949 | 27.47 | -33.27% |
| Test29_17 | 5.141E-05 | 425 | 5662 | 3469 | 9700 | 27.78 | -34.83% |
| Test29_19 | 1.050E-01 | 492 | 7941 | 8170 | 17696 | 15.14 | -65.33% |
| Test29_20 | 1.329E-04 | 532 | 7971 | 7889 | 54621 | 18.85 | -53.39% |
| Test29_22 | 4.850E-05 | 41 | 631 | 364 | 10405 | 45.62 | -31.17% |
| Test29_24 | 3.826E-05 | 2413 | 42786 | 54909 | 337121 | 55.02 | -27.84% |

# Outline

# Summary

Shortcomings of gradient sampling methods to date:

- QP subproblems have potentially many constraints, and
- QP subproblems need to be solved *exactly* in each iteration

Our contributions:

- *inexact* subproblem solves
- *gradient aggregation* to limit subproblem sizes

. . . all while maintaining convergence guarantees of the basic method.

- "Gradient Sampling Methods with Inexact Subproblem Solutions and Gradient Aggregation" https://arxiv.org/abs/2005.07822.