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 Optimization

July 21, 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

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

If f is locally Lipschitz, then ideally one can solve

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

However, this is intractable, so we approximate:

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

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** $\mathcal{G}_k := \{\nabla f(x_k), \nabla f(x_{k,1}), \ldots, \nabla f(x_{k,p})\}$
- **compute** the minimum norm element of $\operatorname{conv}(\mathcal{G}_k)$, 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.

Illustration

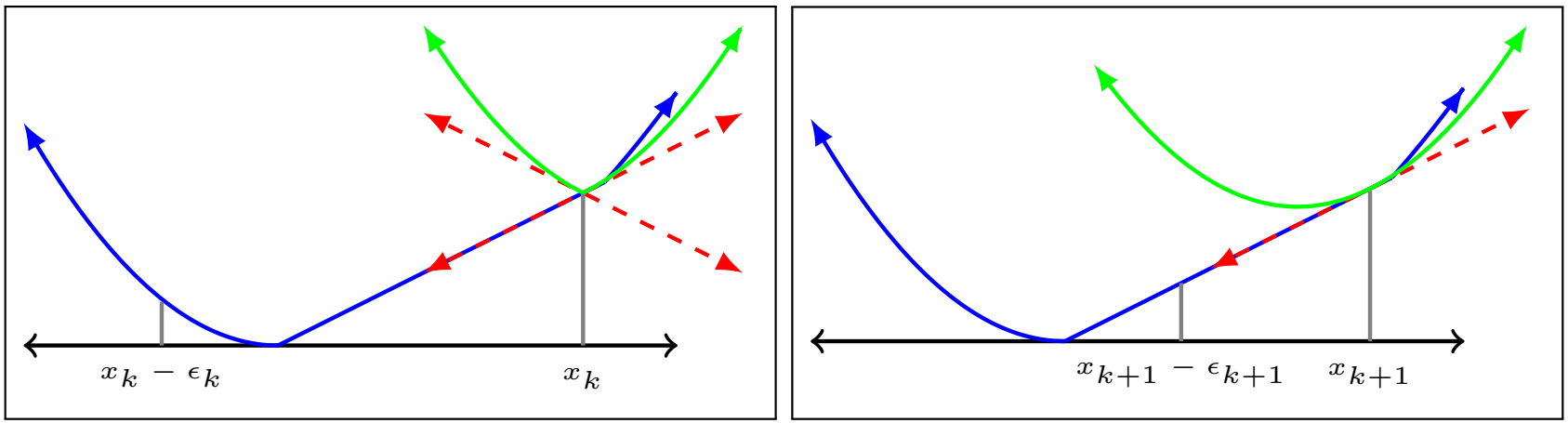


Figure: Illustration of subsequent GS iterations.

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

QP subproblems

Primal-dual form of the gradient sampling QP subproblems:

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

- 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:

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

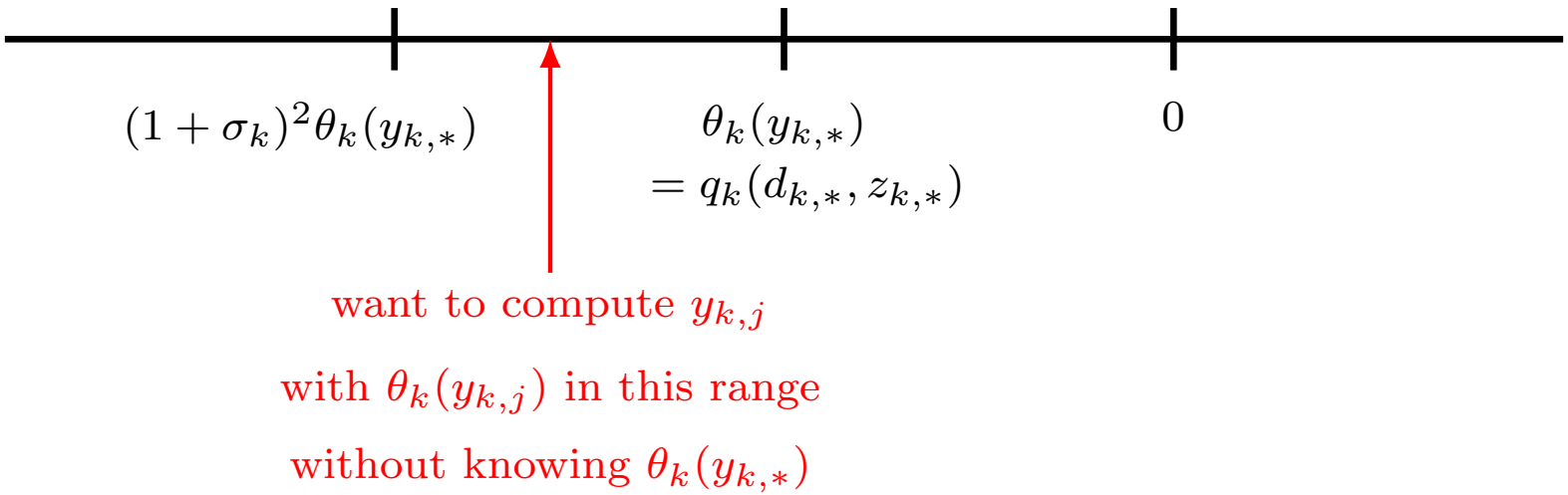
As $y_{k,j} \to y_{k,*}$, this converges to primal-dual solution.

Primal-dual termination test

Consider the primal and dual objective functions:

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

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

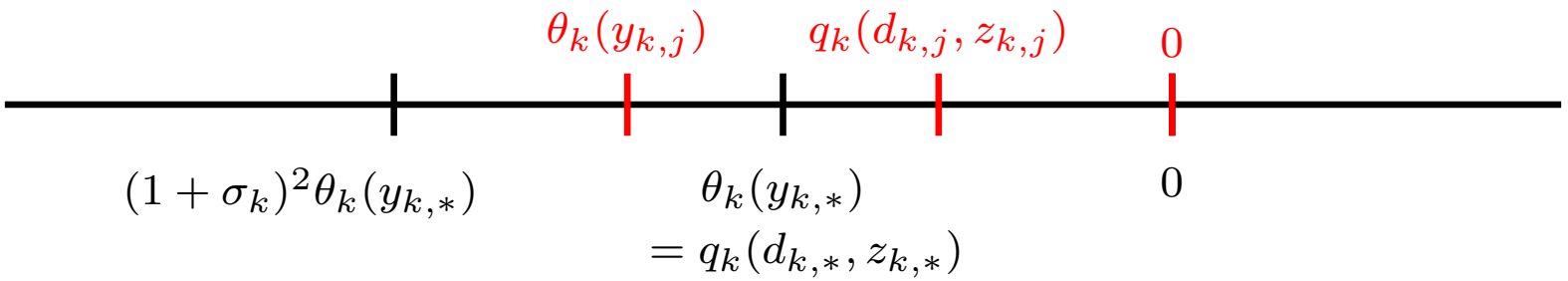


Termination test 1

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.

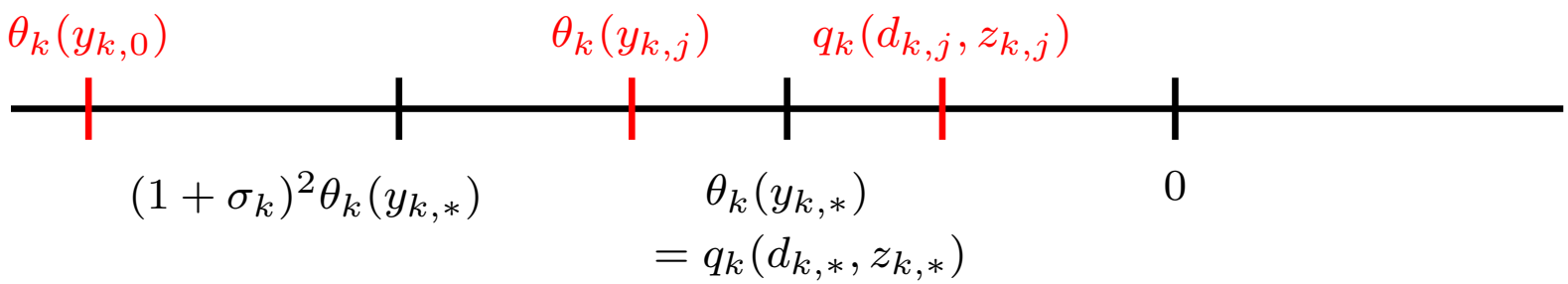


Termination test 2

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

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

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

Common test problems ($n = 1000$); NonOpt limited by time required by LMBM

Table: LMBM versus NonOpt

	LMBM			NonOpt			
name	iters	f-evals	$f(x)$	iters	qp-iters	f-evals	$f(x)$
MaxQ	21940	22808	+4.987830e-06	5443	5447	14283	+9.920069e-05
MxHilb	441	861	+6.166410e-03	142	142	674	+4.648503e-05
Chained_LQ	300	1824	-1.412780e+03	55	283	444	-1.412570e+03
Chained_CB3_1	291	1690	+1.998000e+03	119	272	818	+2.003039e+03
Chained_CB3_2	66	150	+1.998000e+03	91	252	606	+1.998000e+03
ActiveFaces	523	569	+1.376680e-14	15	788	394	+3.961526e-05
Brown_Function_2	493	4217	+2.136910e-09	51	288	405	+9.411906e-02
Chained_Mifflin_2	546	3892	-7.064510e+02	53	305	462	-7.061542e+02
Chained_Crescent_1	177	817	+3.681010e-08	39	60	232	+5.335065e-10
Chained_Crescent_2	903	9626	+1.369240e-04	52	276	456	+2.556368e-01
Test29_2	62	63	+9.815390e-01	1366	1487	8175	+7.677696e-02
Test29_5	1230	4563	+6.434430e-06	125	1563	777	+3.224645e-05
Test29_6	44	48	+2.000000e+00	60	292	478	+2.000982e+00
Test29_11	283	1336	+1.203580e+04	20	293	252	+1.207612e+04
Test29_13	3747	7092	+5.665460e+02	116	1629	1420	+5.700497e+02
Test29_17	962	2247	+3.574260e-03	25	269	271	+1.091778e-03
Test29_19	143	1012	+1.000000e+00	54	310	503	+1.002374e+00
Test29_20	277	3087	+5.000010e-01	83	280	762	+5.000672e-01
Test29_22	21	172	+1.966970e-06	30	279	362	+1.042170e-04
Test29_24	315	1945	+4.232150e-02	44	295	507	+1.099555e-01

https://github.com/frankecurtis/NonOpt

Numerical experiments with NonOpt

Experiments with randomized problems of the form ($n = 500$, $m = 400$):

$$\min_{x \in \mathbb{R}^n} \; g^T x + \tfrac{1}{2} x^T H x + \max_{i \in [m]} (a_i^T x + b_i)$$

Table: Results for `GS-exact` versus `GS-inexact`.

problem #	iters	qp-iters	f-evals	g-evals	objective	% ↓ qp-iters
0	551	2732	4176	29695	+1.239436e-03	
1	562	2661	4262	30286	+1.234747e-03	
2	567	3110	4278	30449	+1.064397e-03	
3	562	3051	4257	30449	+1.084893e-03	
4	616	3109	4669	32627	+1.098173e-03	
5	533	2798	4042	29155	+1.152184e-03	
6	462	2309	3514	25950	+1.201356e-03	
7	498	2213	3753	27147	+1.326716e-03	
8	538	2699	4067	28979	+1.090619e-03	
9	439	2313	3360	25298	+1.241042e-03	
0	548	2614	4136	29387	+1.237740e-03	-4.310055
1	543	2591	4093	29080	+1.239542e-03	-2.616570
2	559	2799	4228	30206	+1.062914e-03	-10.003279
3	554	2812	4195	30081	+1.085707e-03	-7.803190
4	611	3043	4596	32275	+1.099544e-03	-2.149265
5	523	2626	3972	28794	+1.148177e-03	-6.123893
6	460	2243	3483	25312	+1.206694e-03	-2.863402
7	498	2223	3733	26680	+1.328899e-03	+0.445497
8	539	2624	4069	29208	+1.093068e-03	-2.778601
9	441	2277	3381	25389	+1.245562e-03	-1.528737

Outline

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:

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

At the primal-dual optimal solution:

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

Hence, the primal optimal solution is also a solution to

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

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: GS-inexact versus GS-inexact-agg.

problem #	iters	qp iters	f-evals	g-evals	objective	% improvement
0	548	2614	4136	29387	+1.237740e-03	
1	543	2591	4093	29080	+1.239542e-03	
2	559	2799	4228	30206	+1.062914e-03	
3	554	2812	4195	30081	+1.085707e-03	
4	611	3043	4596	32275	+1.099544e-03	
5	523	2626	3972	28794	+1.148177e-03	
6	460	2243	3483	25312	+1.206694e-03	
7	498	2223	3733	26680	+1.328899e-03	
8	539	2624	4069	29208	+1.093068e-03	
9	441	2277	3381	25389	+1.245562e-03	
0	553	2230	4160	29184	+1.238634e-03	-14.701136
1	552	2191	4134	28841	+1.237933e-03	-15.463027
2	571	2455	4259	29752	+1.063100e-03	-12.295404
3	568	2527	4249	29819	+1.082229e-03	-10.142980
4	602	2550	4487	31146	+1.101189e-03	-16.194883
5	523	2175	3906	27390	+1.148150e-03	-17.168026
6	459	1816	3436	24424	+1.201954e-03	-19.057806
7	492	1769	3671	25964	+1.328518e-03	-20.427147
8	537	2128	4004	27967	+1.092298e-03	-18.899768
9	434	1831	3251	23276	+1.245693e-03	-19.598184

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.

ICCOPT 2022

International Conference on Continuous Optimization

July 23–28, 2022