SEGA: Variance Reduction via Gradient Sketching

Filip Hanzely\(^1\)  Konstantin Mishchenko\(^3\)  Peter Richtárik\(^{1,2,3}\)

\(^1\) King Abdullah University of Science and Technology, \(^2\) University of Edinburgh, \(^3\) Moscow Institute of Physics and Technology

Abstract

We propose a randomized first order optimization method—SEGA (SkEtched GrAdient)—which progressively throughout its iterations builds a variance-reduced estimate of the gradient from random linear measurements (sketches) of the gradient. In each iteration, SEGA updates the current estimate of the gradient through a sketch-and-project operation using the information provided by the latest sketch, and this is subsequently used to compute an unbiased estimate of the true gradient through a random relaxation procedure. This unbiased estimate is then used to perform a gradient step. Unlike standard subspace descent methods, such as coordinate descent, SEGA can be used for optimization problems with a non-separable proximal term. We provide a general convergence analysis and prove linear convergence for strongly convex objectives. In the special case of coordinate sketches, SEGA can be enhanced with various techniques such as importance sampling, minibatching and acceleration, and its rate is up to a small constant factor identical to the best-known rate of coordinate descent.

1 Introduction

Consider the optimization problem

\[
\min_{x \in \mathbb{R}^n} F(x) := f(x) + R(x),
\]

where \(f : \mathbb{R}^n \rightarrow \mathbb{R}\) is smooth and \(\mu\)-strongly convex, and \(R : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}\) is a closed convex regularizer. In some applications, \(R\) is either the indicator function of a convex set or a sparsity inducing non-smooth penalty such as \(\ell_1\)-norm. We assume that the proximal operator of \(R\), defined by

\[
\text{prox}_R(x) := \arg\min_{y \in \mathbb{R}^n} \left\{ R(y) + \frac{1}{2\sigma} \|y - x\|_B^2 \right\},
\]

is easily computable (e.g., in closed form).

Above we use the weighted Euclidean norm \(\|x\|_B := \langle x, x \rangle_B^{1/2}\), where \(\langle x, y \rangle_B := \langle Bx, y \rangle\) is a weighted inner product associated with a positive definite weight matrix \(B > 0\). Strong convexity of \(f\) is defined with respect to the same product and norm\(^1\).

1.1 Gradient sketching

In this paper we design proximal gradient-type methods for solving (1) without assuming that the true gradient of \(f\) is available. Instead, we assume that an oracle provides a random linear transformation (i.e., a sketch) of the gradient, which is the information available to drive the iterative

\[1 f\) is \(\mu\)-strongly convex if \(f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle_B + \frac{\mu}{2} \|x - y\|_B^2\) for all \(x, y \in \mathbb{R}^n\).
process. In particular, given a fixed distribution $\mathcal{D}$ over matrices $S \in \mathbb{R}^{n \times b}$ ($b \geq 1$ but does not need to be fixed), and a query point $x \in \mathbb{R}^n$, our oracle provides us the random linear transformation of the gradient given by

$$
\zeta(S, x) \overset{\text{def}}{=} S^T \nabla f(x) \in \mathbb{R}^b, \quad S \sim \mathcal{D}.
$$

(2)

Information of this type is available/used in a variety of scenarios. For instance, randomized coordinate descent (CD) methods use oracle $\overset{(1)}{\zeta}$ with $\mathcal{D}$ corresponding to a distribution over standard basis vectors. Minibatch/parallel variants of CD methods utilize oracle $\overset{(2)}{\zeta}$ with $\mathcal{D}$ corresponding to a distribution over random column submatrices of the identity matrix. If one is prepared to use difference of function values to approximate directional derivatives, one can apply our oracle model to zeroth-order optimization $\overset{(3)}{\zeta}$. Indeed, the directional derivative of $f$ in a random direction $S = s \in \mathbb{R}^{n \times 1}$ can be approximated by $\zeta(s, x) \approx \frac{1}{b} (f(x + \epsilon s) - f(x))$, where $\epsilon > 0$ is sufficiently small.

We now illustrate this concept using two examples.

**Example 1.1 (Sketches).** (i) Coordinate sketch. Let $\mathcal{D}$ be the uniform distribution over standard unit basis vectors $e_1, e_2, \ldots, e_n$ of $\mathbb{R}^n$. Then $\zeta(e_i, x) = e_i^T \nabla f(x)$, i.e., the $i^{th}$ partial derivative of $f$ at $x$. (ii) Gaussian sketch. Let $\mathcal{D}$ be the standard Gaussian distribution in $\mathbb{R}^n$. Then for $s \sim \mathcal{D}$ we have $\zeta(s, x) = s^T \nabla f(x)$, i.e., the directional derivative of $f$ at $x$ in direction $s$.

### 1.2 Related work

In the last decade, stochastic gradient-type methods for solving problem $\overset{(1)}{\zeta}$ have received unprecedented attention by theoreticians and practitioners alike. Specific examples of such methods are stochastic gradient descent (SGD) $\overset{(4)}{\zeta}$, variance-reduced variants of SGD such as SAG $\overset{(10)}{\zeta}$, SAGA $\overset{(12)}{\zeta}$, SVRG $\overset{(22)}{\zeta}$, and their accelerated counterparts $\overset{(46)}{\zeta}$. While these methods are specifically designed for objectives formulated as an expectation or a finite sum, we do not assume such a structure. Moreover, these methods utilize a fundamentally different stochastic gradient information: they have access to an unbiased gradient estimator. In contrast, we do not assume that $\overset{(2)}{\zeta}$ is an unbiased estimator of $\nabla f(x)$. In fact, $\zeta(S, x) \in \mathbb{R}^b$ and $\nabla f(x) \in \mathbb{R}^n$ do not even necessarily belong to the same space. Therefore, our algorithms and results are complementary to the above line of research.

While the gradient sketch $\zeta(S, x)$ does not immediately lead to an unbiased estimator of the gradient, SEGA uses the information provided in the sketch to construct an unbiased estimator of the gradient via a sketch-and-project process. Sketch-and-project iterations were introduced in $\overset{(15)}{\zeta}$ in the context of linear feasibility problems. A dual view uncovering a direct relationship with stochastic subspace ascent methods was developed in $\overset{(16)}{\zeta}$. The latest and most in-depth treatment of sketch-and-project for linear feasibility is based on the idea of stochastic reformulations $\overset{(42)}{\zeta}$. Sketch-and-project can be combined with Polyak $\overset{(29)}{\zeta}$ and Nesterov momentum $\overset{(14)}{\zeta}$, extended to convex feasibility problems $\overset{(30)}{\zeta}$, matrix inversion $\overset{(13)}{\zeta}$ $\overset{(17)}{\zeta}$ $\overset{(14)}{\zeta}$, and empirical risk minimization $\overset{(13)}{\zeta}$ $\overset{(17)}{\zeta}$.

The line of work most closely related to our setup is that on randomized coordinate/subspace descent methods $\overset{(44)}{\zeta}$ $\overset{(16)}{\zeta}$. Indeed, the information available to these methods is compatible with our oracle for specific distributions $\mathcal{D}$. However, the main disadvantage of these methods is that they cannot handle non-separable regularizers $R$. In contrast, the algorithm we propose—SEGA—works for any regularizer $R$. In particular, SEGA can handle non-separable constraints even with coordinate sketches, which is out of range of current CD methods. Hence, our work could be understood as extending the reach of coordinate and subspace descent methods from separable to arbitrary regularizers, which allows for a plethora of new applications. Our method is able to work with an arbitrary regularizer due to its ability to build an unbiased variance-reduced estimate of the gradient of $f$ throughout the iterative process from the random sketches provided by the oracle. Moreover, and unlike coordinate descent, SEGA allows for general sketches from essentially any distribution $\mathcal{D}$. 

---

1. Unlabeled
2. Unlabeled
3. Unlabeled
4. Unlabeled
5. Unlabeled
6. Unlabeled
7. Unlabeled
8. Unlabeled
9. Unlabeled
10. Unlabeled
11. Unlabeled
12. Unlabeled
13. Unlabeled
14. Unlabeled
15. Unlabeled
16. Unlabeled
17. Unlabeled
18. Unlabeled
19. Unlabeled
20. Unlabeled
21. Unlabeled
22. Unlabeled
23. Unlabeled
24. Unlabeled
25. Unlabeled
26. Unlabeled
27. Unlabeled
28. Unlabeled
29. Unlabeled
30. Unlabeled
31. Unlabeled
32. Unlabeled
33. Unlabeled
34. Unlabeled
35. Unlabeled
36. Unlabeled
37. Unlabeled
38. Unlabeled
39. Unlabeled
40. Unlabeled
41. Unlabeled
42. Unlabeled
43. Unlabeled
44. Unlabeled
45. Unlabeled
46. Unlabeled
47. Unlabeled
48. Unlabeled
49. Unlabeled
50. Unlabeled
51. Unlabeled
52. Unlabeled
53. Unlabeled
54. Unlabeled
55. Unlabeled
56. Unlabeled
57. Unlabeled
58. Unlabeled
59. Unlabeled
60. Unlabeled
61. Unlabeled
62. Unlabeled
63. Unlabeled
64. Unlabeled
65. Unlabeled
66. Unlabeled
67. Unlabeled
68. Unlabeled
69. Unlabeled
70. Unlabeled
71. Unlabeled
72. Unlabeled
73. Unlabeled
74. Unlabeled
75. Unlabeled
76. Unlabeled
77. Unlabeled
78. Unlabeled
79. Unlabeled
80. Unlabeled
81. Unlabeled
82. Unlabeled
83. Unlabeled
84. Unlabeled
85. Unlabeled
86. Unlabeled
87. Unlabeled
88. Unlabeled
89. Unlabeled
90. Unlabeled
91. Unlabeled
92. Unlabeled
93. Unlabeled
94. Unlabeled
95. Unlabeled
96. Unlabeled
97. Unlabeled
98. Unlabeled
99. Unlabeled
100. Unlabeled
Another stream of work on designing gradient-type methods without assuming perfect access to the gradient is represented by the *inexact gradient descent* methods \([9, 11, 45]\). However, these methods deal with deterministic estimates of the gradient and are not based on linear transformations of the gradient. Therefore, this second line of research is also significantly different from what we do here.

1.3 Outline

We describe SEGA in Section 2. Convergence results for general sketches are described in Section 3. Refined results for coordinate sketches are presented in Section 4, where we also describe and analyze an accelerated variant of SEGA. Experimental results can be found in Section 5. Conclusions are drawn and potential extensions outlined in Appendix A. Proofs of the main results can be found in Appendices B and C. An aggressive *subspace* variant of SEGA is described and analyzed in Appendix D. A simplified analysis of SEGA in the case of coordinate sketches and for \(R = 0\) is developed in Appendix E (under standard assumptions as in the main paper) and F (under alternative assumptions). Extra experiments for additional insights are included in Appendix G.

Notation. We introduce notation where needed. We also provide a notation table in Appendix H.

2 The SEGA Algorithm

In this section we introduce a learning process for estimating the gradient from the sketched information provided by (2); this will be used as a subroutine of SEGA.

Let \(x^k\) be the current iterate, and let \(h^k\) be the current estimate of the gradient of \(f\). The oracle queried, and we receive new information in the form of the sketched gradient (2). Then, we would like to update \(h^k\) based on the new information. We do this using a *sketch-and-project* process \([15, 16, 42]\): we set 

\[
\begin{align*}
\text{arg min}_{h \in \mathbb{R}^n} & \quad \|h - h^k\|_B^2 \\
\text{subject to} & \quad S_k^T h = S_k^T \nabla f(x^k).
\end{align*}
\]

The closed-form solution of (3) is

\[
\begin{align*}
h^{k+1} &= h^k - B^{-1} Z_k (h^k - \nabla f(x^k)) \\
&= (I - B^{-1} Z_k) h^k + B^{-1} Z_k \nabla f(x^k),
\end{align*}
\]

where \(Z_k \overset{\text{def}}{=} S_k (S_k^T B^{-1} S_k)^{-1} S_k^T\). Notice that \(h^{k+1}\) is a *biased* estimator of \(\nabla f(x^k)\). In order to obtain an unbiased gradient estimator, we introduce a random variable \(\theta_k = \theta(S_k)\) for which

\[
\mathbb{E}_D [\theta_k Z_k] = B.
\]

If \(\theta_k\) satisfies (5), it is straightforward to see that the random vector

\[
\begin{align*}
g^k &\overset{\text{def}}{=} (1 - \theta_k) h^k + \theta_k h^{k+1} = h^k + \theta_k B^{-1} Z_k (\nabla f(x^k) - h^k)
\end{align*}
\]

is an *unbiased estimator* of the gradient:

\[
\mathbb{E}_D [g^k] = \nabla f(x^k).
\]

Finally, we use \(g^k\) instead of the true gradient, and perform a proximal step with respect to \(R\). This leads to a new optimization method, which we call *SkEtched GrAdient Method (SEGA)* and describe in Algorithm 1. We stress again that the method does not need the access to the full gradient.

\footnote{Such a random variable may not exist. Some sufficient conditions are provided later.}
Based on the above observations, we conclude that \( h \) updates all coordinates in every iteration, allows for arbitrary proximal term, random partial derivative of \( f \), which means that \( h \) does not depend on \( B \). This can equivalently be written as
\[
\mathbb{E}_D [h_{k+1}] = \mathbb{E}_D [h_k + e_i^T (\nabla f(x^k) - h_k) e_i],
\]
which can equivalently be written as
\[
h_{k+1} = e_i^T (\nabla f(x^k) - h_k)
\]
for \( j \neq i \). Note that \( h_{k+1} \) does not depend on \( B \). If we choose \( \theta_k = \theta(S_k) = 1/p_i \), then
\[
\mathbb{E}_D [\theta_k Z_k] = \sum_{i=1}^n p_i \frac{1}{p_i} e_i (e_i B^{-1})^{-1} e_i^T = \sum_{i=1}^n \frac{e_i e_i^T}{1/p_i} = B
\]
which means that \( \theta_k \) is a bias-correcting random variable. We then get
\[
g_k = h_k + e_i^T (\nabla f(x^k) - h_k) e_i.
\]

In the setup of Example 2.1, both SEGA and CD obtain new gradient information in the form of a random partial derivative of \( f \). However, the two methods perform a different update: (i) SEGA allows for arbitrary proximal term, CD allows for separable one only \([46, 27, 12]\); (ii) While SEGA updates all coordinates in every iteration, CD updates a single coordinate only; (iii) If we force \( h_k = 0 \) in SEGA and use coordinate sketches, the method transforms into CD.

Based on the above observations, we conclude that SEGA can be applied in more general settings for the price of potentially more expensive iteration.\(^3\) For intuition-building illustration of how SEGA

---

\(^3\) Forming vector \( g \) and computing the prox.
works, Figure 1 shows the evolution of iterates of both SEGA and CD applied to minimizing a simple quadratic function in 2 dimensions. For more figures of this type, including the composite case where CD does not work, see Appendix G.1.

In Section 4 we show that SEGA enjoys, up to a small constant factor, the same theoretical iteration complexity as CD. This remains true when comparing state-of-the-art variants of CD with importance sampling, parallelism/minibatching and acceleration with the corresponding variants of SEGA.

Remark 2.2. Nontrivial sketches $S$ and metric $B$ might, in some applications, bring a substantial speedup against the baseline choices mentioned in Example 2.1. Appendix D provides one example: there are problems where the gradient of $f$ always lies in a particular $d$-dimensional subspace of $\mathbb{R}^n$. In such a case, suitable choice of $S$ and $B$ leads to $\mathcal{O}\left(\frac{n}{d}\right)$--times faster convergence compared to the setup of Example 2.1. In Section 5.3 we numerically demonstrate this claim.

3 Convergence of SEGA for General Sketches

In this section we state a linear convergence result for SEGA (Algorithm 1) for general sketch distributions $D$ under smoothness and strong convexity assumptions.

3.1 Smoothness assumptions

We will use the following general version of smoothness.

Assumption 3.1 (Q-smoothness). Function $f$ is Q-smooth with respect to $B$, where $Q > 0$ and $B > 0$. That is, for all $x, y$, the following inequality is satisfied:

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle_B \geq \frac{1}{2} \| \nabla f(x) - \nabla f(y) \|_Q^2.$$

Assumption 3.1 is not standard in the literature. However, as Lemma B.1 states, in the special case of $B = I$ and $Q = M^{-1}$, it reduces to $M$-smoothness (see Assumption 3.2), which is a common assumption in modern analysis of CD methods.

Assumption 3.2 (M-smoothness). Function $f$ is $M$-smooth for some matrix $M > 0$. That is, for all $x, y$, the following inequality is satisfied:

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2} \| x - y \|_M^2.$$

Assumption 3.2 is fairly standard in the CD literature. It appears naturally in various application such as empirical risk minimization with linear predictors and is a baseline in the development of minibatch CD methods [31, 38, 36, 39]. We will adopt this notion in Section 4 when comparing SEGA to coordinate descent. Until then, let us consider the more general Assumption 3.1.

3.2 Main result

Now we present one of the key theorems of the paper, stating a linear convergence of SEGA.

Theorem 3.3. Assume that $f$ is Q--smooth with respect to $B$, and $\mu$--strongly convex. Fix $x^0, h^0 \in \text{dom}(F)$ and let $x^k, h^k$ be the random iterates produced by SEGA. Choose stepsize $\alpha > 0$ and Lyapunov parameter $\sigma > 0$ so that

$$\alpha \left( 2(\mathbf{C} - B) + \sigma \mu B \right) \leq \sigma \mathbb{E}_D [Z], \quad \alpha \mathbf{C} \leq \frac{1}{2} \left( Q - \sigma \mathbb{E}_D [Z] \right),$$

where $\mathbf{C} \overset{\text{def}}{=} \mathbb{E}_D [\theta_2^2 Z_k]$. Then $\mathbb{E} \left[ \Phi^k \right] \leq (1 - \alpha \mu)^k \Phi^0$ for Lyapunov function $\Phi_k \overset{\text{def}}{=} \| x^k - x^* \|_B^2 + \sigma \alpha \| h^k - \nabla f(x^*) \|_B^2$, where $x^*$ is a solution of (1).
Note that \( \Phi^k \to 0 \) implies \( h^k \to \nabla f(x^*) \). Therefore SEGA is variance reduced, in contrast to CD in the non-separable proximal setup, which does not converge to the solution. If \( \sigma \) is small enough so that \( Q - \sigma E_D[Z] > 0 \), one can always choose stepsize \( \alpha \) satisfying

\[
\alpha \leq \min \left\{ \frac{\lambda_{\min}(E_D[Z])}{\lambda_{\max}(2\sigma^{-1}(C-B)+\mu B)}, \frac{\lambda_{\min}(Q-\sigma E_D[Z])}{2\lambda_{\max}(C)} \right\} \tag{13}
\]

and inequalities (12) will hold. Therefore, we get the next corollary.

**Corollary 3.4.** If \( \sigma < \frac{\lambda_{\min}(Q)}{\lambda_{\max}(E_D[Z])} \), \( \alpha \) satisfies (13) and \( k \geq \frac{1}{\alpha \mu} \log \frac{2\epsilon}{\sqrt{\sigma}} \), then \( \mathbb{E} [||x^k - x^*||_B^2] \leq \epsilon \).

As Theorem 3.3 is rather general, we also provide a simplified version thereof, complete with a simplified analysis (Theorem E.1 in Appendix E). In the simplified version we remove the proximal setting (i.e., we set \( R = 0 \)), assume \( L \)-smoothness and only consider coordinate sketches with uniform probabilities. The result is provided as Corollary 3.5.

**Corollary 3.5.** Let \( B = I \) and choose \( D \) to be the uniform distribution over unit basis vectors in \( \mathbb{R}^n \). If the stepsize satisfies \( 0 < \alpha \leq \min \{(1 - L\sigma/n)/(2Ln), n^{-1} (\mu + 2(n-1)/\sigma)^{-1}\} \), then \( \mathbb{E}_{D} [\Phi^{k+1}] \leq (1 - \alpha \mu)\Phi^k \), and therefore the iteration complexity is \( O(nL/\mu) \).

**Remark 3.6.** In the fully general case, one might choose \( \alpha \) to be bigger than bound (13), which depends on eigen properties of \( E_D[Z] \), \( C, Q, B \), leading to a better overall complexity. However, in the simple case with \( B = I \), \( Q = I \) and \( S_k = e_{i_k} \) with uniform probabilities, bound (13) is tight.

## 4 Convergence of SEGA for Coordinate Sketches

In this section we compare SEGA with coordinate descent. We demonstrate that, specialized to a particular choice of the distribution \( D \) (where \( S \) is a random column submatrix of the identity matrix), which makes SEGA use the same random gradient information as that used in modern randomized CD methods, SEGA attains, up to a small constant factor, the same convergence rate as CD methods.

Firstly, in Section 4.2 we develop SEGA with in a general setup known as arbitrary sampling \([41, 40, 37, 38, 6]\) (Theorem 4.2). Then, in Section 4.3 we develop an accelerated variant of SEGA (see Theorem 4.5) for arbitrary sampling as well. Lastly, Corollary 4.3 and Corollary 4.4 provide us with importance sampling for both nonaccelerated and accelerated method, which matches up to a constant factor cutting-edge CD rates \([41, 3] \) under the same oracle and assumptions\(^3\). Table 1 summarises the results of this section. We provide all proofs for this section in Appendix E.

\(^2\)The standard \( L \)-smoothness assumption is a special case of \( M \)-smoothness for \( M = LI \), and hence is less general than both \( M \)-smoothness and \( Q \)-smoothness with respect to \( B \).

\(^3\)There was recently introduced a notion of importance minibatch sampling for coordinate descent \([20] \). We state, without a proof, that SEGA allows for the same importance sampling as developed in the mentioned paper.

<table>
<thead>
<tr>
<th></th>
<th>CD</th>
<th>SEGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonaccelerated method importance sampling, ( b = 1 )</td>
<td>( \frac{\text{Trace}(M)}{\mu} \log \frac{1}{\epsilon} ) ([13])</td>
<td>( 8.55 \cdot \frac{\text{Trace}(M)}{\mu} \log \frac{1}{\epsilon} )</td>
</tr>
<tr>
<td>Nonaccelerated method arbitrary sampling</td>
<td>( (\max_i \frac{\mu}{p_i \mu}) \log \frac{1}{\epsilon} ) ([41])</td>
<td>( 8.55 \cdot (\max_i \frac{\mu}{p_i \mu}) \log \frac{1}{\epsilon} )</td>
</tr>
<tr>
<td>Accelerated method importance sampling, ( b = 1 )</td>
<td>( 1.62 \cdot \frac{\sum_i \sqrt{M_{ii}}}{\sqrt{\mu}} \log \frac{1}{\epsilon} ) ([3])</td>
<td>( 9.8 \cdot \frac{\sum_i \sqrt{M_{ii}}}{\sqrt{\mu}} \log \frac{1}{\epsilon} )</td>
</tr>
<tr>
<td>Accelerated method arbitrary sampling</td>
<td>( 1.62 \cdot \frac{\sqrt{\max_i \frac{\mu}{p_i \mu}}}{\sqrt{\mu}} \log \frac{1}{\epsilon} ) ([20])</td>
<td>( 9.8 \cdot \sqrt{\max_i \frac{\mu}{p_i \mu}} \log \frac{1}{\epsilon} )</td>
</tr>
</tbody>
</table>

Table 1: Complexity results for coordinate descent (CD) and our sketched gradient method (SEGA), specialized to coordinate sketching, for \( M \)-smooth and \( \mu \)-strongly convex functions.
We now describe the setup and technical assumptions for this section. In order to facilitate a direct comparison with GD (which does not work with non-separable regularizer $R$), for simplicity we consider problem (1) in the simplified setting with $R \equiv 0$. Further, function $f$ is assumed to be $M$–smooth (Assumption 3.2) and $\mu$–strongly convex.

4.1 Defining $\mathcal{D}$: samplings

In order to draw a direct comparison with general variants of GD methods (i.e., with those analyzed in the arbitrary sampling paradigm), we consider sketches in (3) that are column submatrices of the identity matrix: $S = I_S$, where $S$ is a random subset (aka sampling) of $[n] \equiv \{1, 2, \ldots, n\}$. Note that the columns of $I_S$ are the standard basis vectors $e_i$ for $i \in S$ and hence Range ($S$) = Range ($e_i : i \in S$). So, distribution $\mathcal{D}$ from which we draw matrices is uniquely determined by the distribution of sampling $S$. Given a sampling $S$, define $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$ to be the vector satisfying $p_i = \mathbb{P}(e_i \in \text{Range}(S)) = \mathbb{P}(i \in S)$, and $P$ to be the matrix for which $P_{ij} = \mathbb{P} \{ \{i, j\} \subseteq S \}$. Note that $p$ and $P$ are the probability vector and probability matrix of sampling $S$, respectively. We assume throughout the paper that $S$ is proper, i.e., we assume that $p_i > 0$ for all $i$. State-of-the-art minibatch GD methods (including the ones we compare against [31, 20]) utilize large stepsizes related to the so-called ESO Expected Separable Overapproximation (ESO) [38] parameters $v = (v_1, \ldots, v_n)$. ESO parameters play a key role in SEGA as well, and are defined next.

Assumption 4.1 (ESO). There exists a vector $v$ satisfying the following inequality

$$
P \circ M \preceq \text{Diag}(p)\text{Diag}(v),$$

where $\circ$ denotes the Hadamard (i.e., element-wise) product of matrices.

In case of single coordinate sketches, parameters $v$ are equal to coordinate-wise smoothness constants of $f$. An extensive study on how to choose them in general was performed in [38]. For notational brevity, let us set $\hat{P} \overset{\text{def}}{=} \text{Diag}(p)$ and $\hat{V} \overset{\text{def}}{=} \text{Diag}(v)$ throughout this section.

4.2 Non-accelerated method

We now state the convergence rate of (non-accelerated) SEGA for coordinate sketches with arbitrary sampling of subsets of coordinates. The corresponding GD method was developed in [41].

Theorem 4.2. Assume that $f$ is $M$–smooth and $\mu$–strongly convex. Denote $\Psi^k \overset{\text{def}}{=} f(x^k) - f(x^*) + \sigma \|h^k\|_{\hat{P}^{-1}}^2$. Choose $\alpha, \sigma > 0$ such that

$$
\sigma I - \alpha^2 (\hat{V} \hat{P}^{-1} - M) \succeq \gamma \mu \sigma \hat{P}^{-1},
$$

where $\gamma \overset{\text{def}}{=} \alpha - \alpha^2 \max_i \{\frac{v_i}{p_i}\} - \sigma$. Then the iterates of SEGA satisfy $\mathbb{E} [\Psi^k] \leq (1 - \gamma \mu)^k \Psi^0$.

We now give an importance sampling result for a coordinate version of SEGA. We recover, up to a constant factor, the same convergence rate as standard GD [24]. The probabilities we chose are optimal in our analysis and are proportional to the diagonal elements of matrix $M$.

Corollary 4.3. Assume that $f$ is $M$–smooth and $\mu$–strongly convex. Suppose that $\mathcal{D}$ is such that at each iteration standard basis vector $e_i$ is sampled with probability $p_i \propto M_{ii}$. If we choose $\alpha = \frac{0.232}{\text{Trace}(M)}$, $\sigma = \frac{0.061}{\text{Trace}(M)}$, then $\mathbb{E} [\Psi^k] \leq \left(1 - \frac{0.117 \mu}{\text{Trace}(M)}\right)^k \Psi^0$.

The iteration complexities from Theorem 4.2 and Corollary 4.3 are summarized in Table 1. We also state that $\alpha, \sigma$ can be chosen so that (15) holds, and the rate from Theorem 4.2 coincides with the rate from Table 1. Theorem 4.2 and Corollary 4.3 hold even under a non-convex relaxation of strong convexity – Polyak-Łojasiewicz inequality: $\mu (f(x) - f(x^*)) \leq \frac{1}{2} \|\nabla f(x)\|_2^2$. Thus, SEGA works for a certain class of non-convex problems. For an overview on relaxations of strong convexity, see [25].
4.3 Accelerated method

In this section, we propose an accelerated (in the sense of Nesterov’s method [31, 32]) version of SEGA, which we call ASEGA. The analogous accelerated CD method, in which a single coordinate is sampled in every iteration, was developed and analyzed in [3]. The general variant utilizing arbitrary sampling was developed and analyzed in [20].

Algorithm 2: ASEGA: Accelerated SEGA

1. Initialize: $x^0 = y^0 = z^0 \in \mathbb{R}^n$; $h^0 \in \mathbb{R}^n$; $S$; parameters $\alpha, \beta, \tau, \mu > 0$
2. for $k = 1, 2, \ldots$ do
3. 
4. Sample $S_k = I_{S_k}$, where $S_k \sim S$, and compute $g^k, h^{k+1}$ according to (4), (6)
5. $z^k = \frac{1}{1 + \beta \mu}(z^k + \beta \mu x^k - \beta g^k)$

The method and analysis is inspired by [2]. Due to space limitations and technicality of the content, we state the main theorem of this section in Appendix C.4. Here, we provide Corollary 4.4, which shows that Algorithm 2 with single coordinate sampling enjoys, up to a constant factor, the same convergence rate as state-of-the-art accelerated coordinate descent method NUACDM [3].

Corollary 4.4. Let the sampling be defined as follows: $S = \{i\}$ w. p. $p_i \propto \sqrt{M_{ii}}$, for $i \in [n]$. Then there exist acceleration parameters and a Lyapunov function $\Upsilon_k$ such that $f(y^k) - f(x^*) \leq \Upsilon_k$ and $
abla \Upsilon_k \leq (1 - \tau)^k \Upsilon_0 = (1 - O(\sqrt{\beta} / \sum_i \sqrt{M_{ii}})) k \Upsilon_0$.

The iteration complexity provided by Theorem C.5 and Corollary 4.4 are summarized in Table 1.

5 Experiments

In this section we perform numerical experiments to illustrate the potential of SEGA. Firstly, in Section 5.1, we compare it to projected gradient descent (PGD) algorithm. Then in Section 5.2, we study the performance of zeroth-order SEGA (when sketched gradients are being estimated through function value evaluations) and compare it to the analogous zeroth-order method. Lastly, in Section 5.3, we verify the claim from Remark 3.6 that in some applications, particular sketches and metric might lead to a significantly faster convergence. In the experiments where theory-supported stepsizes were used, we obtained them by precomputing strong convexity and smoothness measures.

5.1 Comparison to projected gradient

In this experiment, we show the potential superiority of our method to PGD. We consider the $\ell_2$ ball constrained problem ($R$ is the indicator function of the unit ball) with the oracle providing the sketched gradient in the random Gaussian direction. As we mentioned, a method moving in the gradient direction (analogue of CD), will not converge due as $R$ is not separable. Therefore, we can only compare against the projected gradient. In order to obtain the full gradient for PGD, one needs to gather $n$ sketched gradients and solve a corresponding linear system. As for $f$, we choose 4 different quadratics, see Table 2 (appendix). We stress that these are synthetic problems generated for the purpose of illustrating the potential of our method against a natural baseline. Figure 3 compares SEGA and PGD under various relative cost scenarios of solving the linear system compared to the cost of the oracle calls. The results show that SEGA significantly outperforms PGD as soon as solving the linear system is expensive, and is as fast as PGD even if solving the linear system comes for free.
5.2 Comparison to zeroth-order optimization methods

In this section, we compare SEGA to the random direct search (RDS) method under a zeroth-order oracle and $R = 0$. For SEGA, we estimate the sketched gradient using finite differences. Note that RDS is a randomized version of the classical direct search method [21, 24, 25]. At iteration $k$, RDS moves to $\text{argmin} (f(x^k + \alpha^k s^k), f(x^k - \alpha^k s^k), f(x^k))$ for a random direction $s^k \sim \mathcal{D}$ and a suitable stepsize $\alpha^k$. For illustration, we choose $f$ to be a quadratic problem based on Table 2 and compare both Gaussian and coordinate sketches. Figure 3 shows that SEGA outperforms RDS.

5.3 Subspace SEGA: a more aggressive approach

As mentioned in Remark 3.6, well designed sketches are capable of exploiting structure of $f$ and lead to a better rate. We address this in detail in Appendix D where we develop and analyze a subspace variant of SEGA. To illustrate this phenomenon in a simple setting, we perform experiments for problem (11) with $f(x) = \|Ax - b\|^2$, where $b \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d \times n}$ has orthogonal rows, and with $R$ being the indicator function of the unit ball in $\mathbb{R}^n$. We assume that $n \gg d$. We compare two methods: naiveSEGA, which uses coordinate sketches, and subspaceSEGA, where sketches are chosen as rows of $A$. Figure 4 indicates that subspaceSEGA outperforms naiveSEGA roughly by the factor $\frac{n}{d}$, as claimed in Appendix D.

Figure 4: Comparison of SEGA with sketches from a correct subspace versus coordinate sketches naiveSEGA. Step size chosen according to theory. 1000 dimensional problem.
References


