Revisiting Incremental Gradient Methods

Yuan Gao Sebastian Stich

May 12, 2026

Preface

SGD: The gradients at each iterations are sampled i.i.d. The current randomness does not depend on the previous iterations.

Incremental: The gradients are accessed in a fixed order. The algorithm runs in epochs.

Motivation

- SGD theory is well understood, but deviates from practice.
- Incremental methods are widely used in practice, but its theory is much less developed.
  - Unsatisfactory algorithms for composite optimization problems.
  - (Most) convergence analysis generally follows an epoch-wise pattern, resulting in epoch-size dependent convergence rates.

Problem Formuation

$$\mathbf{x}^\star = \arg\min F(\mathbf{x}) := f(\mathbf{x}) + \psi(\mathbf{x}), \quad f(\mathbf{x}) = \frac{1}{n}\sum_{i=0}^{n-1} f_i(\mathbf{x})$$

f, f_i are convex, and ψ is convex, closed, and proper over the convex domain $\operatorname{dom}\psi$.

Each f_i is L-smooth.

For any $\mathbf{x} \in \operatorname{dom}\psi$, there exists σ^2 such that $\frac{1}{n}\sum_{i=0}^{n-1} \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \sigma^2$.

Algorithm: Proximal Version

Algorithm Proximal Incremental Gradient

Input: $\mathbf{x}_0$ and $\gamma \in \mathbb{R}_+$.
for $k = 0, 1, \ldots, K-1$ **do**
  $\mathbf{x}_k^0 := \mathbf{x}^k$
  for $i = 0, 1, \ldots, n-1$ **do**
    Obtain $\mathbf{g}_k^i := \nabla f_i(\mathbf{x}_k^i)$.
    $\mathbf{x}_k^{i+1} := \mathbf{x}_k^i - \frac{1}{\gamma}\mathbf{g}_k^i$
  $\mathbf{x}^{k+1} := \arg\min_{\mathbf{x} \in \operatorname{dom}\psi} \{\gamma \left\|\mathbf{x} - \mathbf{x}_k^n\right\|^2 + n\psi(\mathbf{x})\}$

Algorithm: Issues

- The proximal operation is only performed once at the end of each epoch. Inner iterations are just gradient steps.
- Problem: the inner iterates $\mathbf{x}_k^i$ are not necessarily in $\operatorname{dom} \psi$. Gradient evaluations might well be undefined.
- Causes: such an algorithm design is tied to the epoch-wise analysis.

Algorithm: Dual Averaging

Algorithm Incremental Dual Averaging

Input: $\mathbf{x}_0, \{\gamma \in \mathbb{R}_+\}_{t=0,\ldots,\infty}$.
for $t = 0, 1, \ldots$ **do**
  Obtain $\mathbf{g}_t = \nabla f_{t \mod n}(\mathbf{x}_t)$.
  $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \operatorname{dom} \psi} \left[\sum_{s=0}^{t} \left(\langle \mathbf{g}_s, \mathbf{x} \rangle + \psi(\mathbf{x}) \right) + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{x}_0\|^2 \right]$

We can also write it in the epoch-wise form. Let's denote $\mathbf{x}^k := \mathbf{x}_{kn}, \mathbf{x}_k^i = \mathbf{x}_{kn+i}$, and $\mathbf{g}_k^i := \nabla f_i(\mathbf{x}_k^i)$.

Epoch-wise Analysis: Sketch I

- We analyze the descent with respect to $\mathbf{x}^k$;
- We effectively treat $\sum_{i=0}^{n-1} \mathbf{g}_{kn+i}$ as a single gradient estimator for $\nabla f(\mathbf{x}^k)$;
- Even in the extreme case that $f_i \equiv f, \forall i \in [n-1]$, this forces $\gamma \geq \mathcal{O}(nL)$.
- This is the classic approach [Mishchenko et al., 2022, 2020, Liu and Zhou, 2024, Josz et al., 2024], but results in epoch-size dependent convergence rates.

Epoch-wise Analysis: Sketch II

Theorem

$$
\begin{aligned}
& n \sum_{k=0}^{K-1} \left(F(\mathbf{x}^{k+1}) - F^\star \right) + \frac{\gamma - nL}{2} \sum_{k=0}^{K-1} \left\| \mathbf{x}^{k+1} - \mathbf{x}^k \right\|^2 \\
\leq & \frac{\gamma R_0^2}{2} - \frac{\gamma R_K^2}{2} + \sum_{k=0}^{K-1} \sum_{i=0}^{n-1} \left(\langle \mathbf{g}_k^i - \nabla f(\mathbf{x}^k), \mathbf{x}^\star - \mathbf{x}^{k+1} \rangle - \beta_f(\mathbf{x}^k, \mathbf{x}^\star) \right),
\end{aligned}
\tag{1}
$$

where we write $R_0 := \|\mathbf{x}_0 - \mathbf{x}^\star\|$ *and* $R_K := \left\|\mathbf{x}^K - \mathbf{x}^\star\right\|$.

Epoch-wise Analysis: Sketch III

Lemma

$$\begin{aligned}&\sum_{i=0}^{n-1}\left(\langle \mathbf{g}_k^i - \nabla f(\mathbf{x}^k), \mathbf{x}^\star - \mathbf{x}^{k+1}\rangle - \beta_f(\mathbf{x}^k, \mathbf{x}^\star)\right)\\ \leq &\frac{Ln}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2 + \frac{3L}{2}\sum_{i=0}^{n-1}\left\|\mathbf{x}_k^i - \mathbf{x}^k\right\|^2.\end{aligned} \tag{2}$$

And when $\gamma \geq 8nL$

$$\sum_{i=0}^{n-1}\left\|\mathbf{x}_k^i - \mathbf{x}^k\right\|^2 \leq \frac{4n^2}{\gamma}\left(F(\mathbf{x}^k) - F^\star\right) + \frac{4n^3\sigma^2}{\gamma^2} \tag{3}$$

Epoch-wise Analysis: Sketch IV

- The stochastic noise term comes with a $\frac{1}{\gamma^2}$ factor, because it is introduced through the drifts $\sum_{i=0}^{n-1} \left\| \mathbf{x}_k^i - \mathbf{x}^k \right\|^2$. This gives us a better asymptotic rate than that of SGD.
- However, the $\gamma \geq 8nL$ requirement gives unsatisfactory optimization term.

Theorem

Given certain choices of γ, it takes at most:

$$K = \frac{24LR_0^2 + 48\sqrt{LR_0^2(F(\mathbf{x}_0) - F^\star)}}{\varepsilon} + \frac{64R_0^2\sqrt{L\sigma}}{\varepsilon^{3/2}}, \tag{4}$$

epochs of Algorithm 2 to find an ε-optimal solution.

Epoch-wise Analysis: Sketch V

- The optimization term can be treated as $\mathcal{O}(\frac{LR_0^2}{\varepsilon})$;
- The asymptotically dominant term is $\mathcal{O}(\frac{1}{\varepsilon^{3/2}})$, better than that of SGD, which is $\mathcal{O}(\frac{1}{\varepsilon^2})$;
- However, in terms of the number of *oracle calls/iterations*, the optimization term becomes $\mathcal{O}(\frac{nLR_0^2}{\varepsilon})$;
- So the rate is WORSE than that of full-batch gradient descent!

Epoch-wise Analysis: Some Comments I

- This is the default analysis template;
- Afaik, there is only one work by Koloskova et al. [2024] that tries to address this issue with the optimization term;
- Koloskova et al. [2024] employed a complicated restarted variant of the virtual iteration techniques that effectively reduces the epoch-wise analysis to a cycle-wise analysis where the cycle size is not directly affected by n;

Epoch-wise Analysis: Some Comments II

- Koloskova et al. [2024]'s restarted virtual iteration technique so far only works for non-composite setting. In fact, the pure virtual iteration technique (without restart) is not even known to work for the composite setting;
- Cycle-wise analysis creates cross-epoch dependences, so the analysis does not work for reshuffling methods, which intuitively should have better rate than incremental;

Question

Is there some analysis that:

- achieves $\mathcal{O}(\frac{LR_0^2}{\varepsilon})$ optimization term;
- achieves $\mathcal{O}(\frac{1}{\varepsilon^{3/2}})$ asymptotic term;
- works for composite optimization problems;
- works for reshuffling methods;
- and is hopefully simpler than the restarted virtual iteration technique?

A Partial Solution: Iteration-wise Analysis

- We can analyze the descent with respect to all iterations $\mathbf{x}_t$;
- This allows us to treat $\mathbf{g}_t$ as a single gradient estimator for $\nabla f(\mathbf{x}_t)$, and thus we can use a small stepsize $\gamma \geq L$.
- For now, it gives us a worse asymptotic term.

Iteration-wise Analysis: Sketch I

Consider the following reformulation of the subproblem:

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}\in\operatorname{dom}\psi}{\arg\min}\, \Phi_t(\mathbf{x}) = \sum_{i=0}^{t} \left(f(\mathbf{x}^\star) + \langle \mathbf{g}_i, \mathbf{x} - \mathbf{x}^\star\rangle + \psi(\mathbf{x})\right) + \frac{\gamma}{2}\left\|\mathbf{x} - \mathbf{x}_0\right\|^2$$

So we directly add $\mathbf{x}^\star$ as the anchor point. Let's write $\Phi_t^\star = \Phi_t(\mathbf{x}_{t+1})$.

Iteration-wise Analysis: Sketch II

Proposition

The iterates of Algorithm 2 satisfy

$$\sum_{t=0}^{T-1} \left(F(\mathbf{x}_{t+1}) - F^\star\right) + \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2$$

$$\leq \frac{\gamma}{2}\left(R_0^2 - R_T^2\right) + \sum_{t=0}^{T-1} \beta_f(\mathbf{x}^\star, \mathbf{x}_{t+1}) + \sum_{t=0}^{T-1} \langle \mathbf{g}_t - \nabla f(\mathbf{x}^\star), \mathbf{x}^\star - \mathbf{x}_{t+1} \rangle.$$

This is just the classical dual averaging analysis, with the slight difference that we use $\mathbf{x}^\star$ as the anchor point instead of $\mathbf{x}_t$.

Iteration-wise Analysis: Sketch III

Let's write $r_t := \beta_f(\mathbf{x}^\star, \mathbf{x}_{t+1}) + \langle \mathbf{g}_t - \nabla f(\mathbf{x}^\star), \mathbf{x}^\star - \mathbf{x}_{t+1} \rangle$.

Lemma

For $T = Kn$ number of iterations, i.e. K epochs, and any $\eta > 0$, we have

$$\sum_{t=0}^{T-1} r_t \leq \frac{n(n+1)\sigma^2}{4\eta} T + \frac{\eta + L}{2} \sum_{t=0}^{T-1} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 . \tag{5}$$

η can be at most $\lesssim \gamma$, otherwise the $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2$ term cannot be absorbed $\implies$ the variance term is only controled by $\frac{1}{\gamma}$, worse than before.

Residual Upper bound I

Let's write $i_t = t \mod n$. We have

$$r_t = f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) + \langle \mathbf{g}_t, \mathbf{x}^\star - \mathbf{x}_{t+1} \rangle = f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) + \langle \nabla f_{i_t}(\mathbf{x}_t), \mathbf{x}^\star - \mathbf{x}_{t+1} \rangle.$$

Add and subtract $f_{i_t}(\mathbf{x}_{t+1})$, $f_{i_t}(\mathbf{x}_t)$, and $f_{i_t}(\mathbf{x}^\star)$:

$$\begin{aligned} r_t = &\big(f(\mathbf{x}_{t+1}) - f_{i_t}(\mathbf{x}_{t+1})\big) + \big(f_{i_t}(\mathbf{x}^\star) - f(\mathbf{x}^\star)\big) \\ &+ \big(f_{i_t}(\mathbf{x}_{t+1}) - f_{i_t}(\mathbf{x}_t) - \langle \nabla f_{i_t}(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle\big) \\ &- \big(f_{i_t}(\mathbf{x}^\star) - f_{i_t}(\mathbf{x}_t) - \langle \nabla f_{i_t}(\mathbf{x}_t), \mathbf{x}^\star - \mathbf{x}_t \rangle\big), \end{aligned}$$

Residual Upper bound II

Summing over one epoch $t = kn + i$ with $i = 0, \ldots, n-1$ gives

$$\sum_{i=0}^{n-1} r_{kn+i} = \sum_{i=0}^{n-1} \left(f(\mathbf{x}_{kn+i+1}) - f_i(\mathbf{x}_{kn+i+1}) \right) + \sum_{i=0}^{n-1} \left(f_i(\mathbf{x}^\star) - f(\mathbf{x}^\star) \right) + \sum_{i=0}^{n-1} \left(\beta_{f_i}(\mathbf{x}_{kn+i}, \mathbf{x}_{kn+i+1}) - \beta_{f_i}(\mathbf{x}_{kn+i}, \mathbf{x}^\star) \right).$$

Since each index appears exactly once in a epoch,

$$\sum_{i=0}^{n-1} \left(f_i(\mathbf{x}^\star) - f(\mathbf{x}^\star) \right) = 0.$$

Residual Upper bound III

Hence

$$\sum_{i=0}^{n-1} r_{kn+i} = \sum_{i=0}^{n-1} \left(f(\mathbf{x}_{kn+i+1}) - f_i(\mathbf{x}_{kn+i+1}) \right) + \sum_{i=0}^{n-1} \left(\beta_{f_i}(\mathbf{x}_{kn+i}, \mathbf{x}_{kn+i+1}) - \beta_{f_i}(\mathbf{x}_{kn+i}, \mathbf{x}^\star) \right).$$

Upper bounding $\beta_{f_i}(\mathbf{x}_{kn+i}, \mathbf{x}_{kn+i+1}) - \beta_{f_i}(\mathbf{x}_{kn+i}, \mathbf{x}^\star)$ is easy, we use smoothness and convexity:

$$\beta_{f_i}(\mathbf{x}_{kn+i}, \mathbf{x}_{kn+i+1}) - \beta_{f_i}(\mathbf{x}_{kn+i}, \mathbf{x}^\star) \leq \frac{L}{2} \left\| \mathbf{x}_{kn+i+1} - \mathbf{x}_{kn+i} \right\|^2.$$

Bounding the functional differences is more tricky, and I'm not so sure if there is a prettier way to do it.

Residual Upper bound IV

Let's write $q_i(\mathbf{x}) := f(\mathbf{x}) - f_i(\mathbf{x})$, and $\xi_i(\mathbf{x}) := \nabla f(\mathbf{x}) - \nabla f_i(\mathbf{x})$. And we write $\Delta_k := \sum_{i=0}^{n-1} f(\mathbf{x}_{kn+i+1} - f_i(\mathbf{x}_{kn+i+1})) = \sum_{i=0}^{n-1} q_i(\mathbf{x}_{kn+i+1})$.
Note that $\sum_{i=0}^{n-1} q_i(\mathbf{x}) = 0$ for any $\mathbf{x}$. So:

$$\begin{aligned}\Delta_k &= \sum_{i=0}^{n-1} \left(q_i(\mathbf{x}_{kn+i+1}) - q_i(\mathbf{x}_{kn})\right) \\ &= \sum_{i=0}^{n-1} \sum_{j=0}^{i} \left(q_i(\mathbf{x}_{kn+j+1}) - q_i(\mathbf{x}_{kn+j})\right)\end{aligned}$$

Residual Upper bound V

Write $d_{k,j} = \mathbf{x}_{kn+j+1} - \mathbf{x}_{kn+j}$, by FTC:

$$\begin{aligned}
\Delta_k &= \sum_{i=0}^{n-1} \sum_{j=0}^{i} \int_0^1 \langle \xi_i(\mathbf{x}_{kn+j} + \tau(d_{k,j})), d_{k,j} \rangle d\tau \\
&= \sum_{j=0}^{n-1} \int_0^1 \sum_{i=j}^{n-1} \langle \xi_i(\mathbf{x}_{kn+j} + \tau(d_{k,j})), d_{k,j} \rangle d\tau \\
&= \sum_{j=0}^{n-1} \int_0^1 \langle S_{k,j}, d_{k,j} \rangle d\tau,
\end{aligned}$$

Residual Upper bound VI

where we write $S_{k,j} := \sum_{i=j}^{n-1} \xi_i(\mathbf{x}_{kn+j} + \tau(d_{k,j}))$. Now for any $\eta > 0$, we apply Young's inequality:

$$\Delta_k \leq \sum_{j=0}^{n-1} \int_0^1 \left(\frac{\|S_{k,j}\|^2}{2\eta} + \frac{\eta}{2} \|d_{k,j}\|^2 \right) d\tau$$

$$= \frac{\eta}{2} \sum_{j=0}^{n-1} \|d_{k,j}\|^2 + \frac{1}{2\eta} \sum_{j=0}^{n-1} \int_0^1 \|S_{k,j}\|^2 \, d\tau$$

We also have:

$$\|S_{k,j}\|^2 \leq (n-j) \sum_{i=j}^{n-1} \|\xi_i(\mathbf{x}_{kn+j} + \tau(d_{k,j}))\|^2 \leq n(n-j)\sigma^2$$

Residual Upper bound VII

So

$$\Delta_k \leq \frac{\eta}{2} \sum_{j=0}^{n-1} \left\| d_{k,j} \right\|^2 + \frac{n(n+1)\sigma^2}{2\eta}.$$

Putting these together:

$$\sum_{i=0}^{n-1} r_{kn+i} \leq \frac{n(n+1)\sigma^2}{2\eta} + \frac{L+\eta}{2} \sum_{i=0}^{n-1} d_{k,i}.$$

Summing k up, and assuming that $T = Kn$ for some K, we get the desired result.

Iteration-wise Analysis: Final Rate

Theorem

Given $\gamma = \max\{2L, \sqrt{\frac{2n^2\sigma^2 T}{R_0^2}}\}$, *it takes at most*

$$T = \frac{2LR_0^2}{\varepsilon} + \frac{2n^2\sigma^2 R_0^2}{\varepsilon^2}$$

iterations/oracle calls of Algorithm 2 to find an ε*-optimal solution.*

Iteration-wise Analysis: Some Comments

- The optimization term is $\mathcal{O}(\frac{LR_0^2}{\varepsilon})$, which is the same as that of SGD, and is independent of the epoch size n;
- The asymptotic term is $\mathcal{O}(\frac{n^2\sigma^2R_0^2}{\varepsilon^2})$, which is n^2 times worse than that of SGD;
- The bottleneck comes from our residual upper bound. We have to upper bound the residual using the variance assumption directly, instead of going through some squared primal distances $\implies$ we only have the $\frac{1}{\gamma}$ factor, instead of the $\frac{1}{\gamma^2}$ factor;
- Directly concatenating the epoch-wise analysis with the iteration-wise analysis does not seem to work.

Summary

- Epoch-wise analysis gives us good asymptotic term, but bad optimization term;
- Iteration-wise analysis gives us good optimization term, but bad asymptotic term;
- I have no idea how to get the best of both worlds. Please let me know if you have any ideas!!

References I

Cedric Josz, Lexiao Lai, and Xiaopeng Li. Proximal random reshuffling under local lipschitz continuity. *arXiv preprint arXiv:2408.07182*, 2024.

Anastasia Koloskova, Nikita Doikov, Sebastian U Stich, and Martin Jaggi. On convergence of incremental gradient for non-convex smooth functions. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ZRMQX6aTUS.

Zijian Liu and Zhengyuan Zhou. On the last-iterate convergence of shuffling gradient methods. In *International Conference on Machine Learning*, pages 32471–32508. PMLR, 2024.

References II

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33: 17309–17320, 2020.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Proximal and federated random reshuffling. In *International Conference on Machine Learning*, pages 15718–15749. PMLR, 2022.