A unified variance-reduced accelerated gradient method for convex optimization

Guanghui Lan, Zhize Li, Yi Zhou

Outline

Convex finite-sum optimization

Randomized incremental gradient methods

Our algorithm *Varag*

- Convergence results
- Numerical experiments

Future works

Problem of interest: convex finite-sum optimization

?

$$\psi^* := \min_{x \in X} \left\{ \psi(x) := \tfrac{1}{m} \textstyle\sum_{i=1}^m f_i(x) + h(x) \right\}.$$

- Smooth and convex with L_i-Lipschitz continuous gradient over X
- Simple but possibly nonsmooth over X

Let $f(x) := \frac{1}{m} \sum_{i=1}^m f_i(x)$, we assume that f is **possibly strongly convex** with modulus $\mu \geq 0$.

Problem of interest: convex finite-sum optimization

?

$$\psi^* := \min_{x \in X} \left\{ \psi(x) := \frac{1}{m} \sum_{i=1}^m f_i(x) + h(x) \right\}.$$

- Wide range of applications in machine learning, statistical inference and image processing.
- Take l_2-regularized logistic regression problem as an example

$$f_i(x) = l_i(x) := \frac{1}{N_i} \sum_{j=1}^{N_i} \log(1 + \exp(-b_j^i {a_j^i}^T x)), \ i = 1, \ldots, m, \quad w(x) = R(x) := \frac{1}{2}\|x\|_2^2,$$

  - f_i is the loss function based on training data $\{a_j^i, b_j^i\}_j^{N_i}$, or the loss function associated with agent i for a distributed optimization problem.
  - Minimization of the empirical risk

$$f_i(x) = l_i(x) := \mathbb{E}_{\xi_i}[\log(1 + \exp(-\xi_i^T x))], \ i = 1, \ldots, m,$$

  - f_i given in the form of expectation where ξ_i models the underlying distribution for training dataset i (of agent i for a distributed problem)
  - Minimization of the generalized risk

Randomized incremental gradient (RIG) methods

- Derived from SGD and the idea of reducing variance of the gradient estimator
- SVRG[JZ13] exhibits linear rate of convergence $\mathcal{O}\{(m + {}^{L}/_{\mu})\log({}^{1}/_{\epsilon})\}$, same result for Prox-SVRG[XZ14], SAGA[DBL14] and SARAH[NLST17] for **strongly convex problems**
  - Update exact gradient $\tilde{g}$ at the outer loop and a gradient of the component function in the inner loop
  - Variance of G_t vanishes as algorithm proceeds
- SVRG++[AY16] obtains $\mathcal{O}\{m\log({}^{1}/_{\epsilon}) + {}^{L}/_{\epsilon}\}$ for smooth convex problems

They are NOT optimal RIG methods!

$$x_t = x_{t-1} - \eta G_t$$

$$\tilde{g} = \nabla f(\tilde{x})$$

$$G_t = \nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(\tilde{x}) + \tilde{g}$$

$$y_i^t = \begin{cases} \nabla f_i(x^t), & i = i_t, \\ y_i^{t-1}, & \text{otherwise}, \end{cases}$$

$$G_t = \nabla f_i(x_t) - y_i^{t-1} + \frac{1}{m}\sum_{i=1}^{m} y_i^{t-1}$$

$$G_0 = \nabla f(\tilde{x})$$

$$G_t = \nabla f_{i_t}(x_{t-1}) - \nabla f_{i_t}(x_{t-2}) + G_{t-1}$$

Optimal RIG methods

Accelerated RIG methods: Catalyst[LMH15], RPGD[LZ17], RGEM[LZ18], and Katyusha[A17], etc.

- All exhibit $\mathcal{O}\{\left(m + \sqrt{mL/\mu}\right)\log(1/\epsilon)\}$ for strongly convex problems
- Except Katyusha^{ns}[A17], none of these methods can be used directly to solve smooth convex problems. They required perturbation technique. Katyusha^{ns} is not advantageous over accelerated gradient method.
- Except RGEM[LZ18], none of the optimal methods can solve stochastic finite-sum problems
- They are assume the strongly convexity comes from regularizer term $h(x)$
- None of them are unified methods that can be adjust to ill-conditioned problem, e.g., μ is very small.

$$\psi^* := \min_{x \in X} \left\{ \psi(x) := \tfrac{1}{m} \textstyle\sum_{i=1}^m f_i(x) + h(x) \right\}.$$

Table 1: Summary of the recent results on accelerated RIG methods

Algorithms	Deterministic smooth strongly convex	Deterministic smooth convex
RPDG[18]	$\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\mu}}) \log \frac{1}{\epsilon}\right\}$	$\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\epsilon}}) \log \frac{1}{\epsilon}\right\}^1$
Catalyst[20]	$\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\mu}}) \log \frac{1}{\epsilon}\right\}^1$	$\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\epsilon}}) \log^2 \frac{1}{\epsilon}\right\}^1$
Katyusha[1]	$\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\mu}}) \log \frac{1}{\epsilon}\right\}$	$\mathcal{O}\left\{(m \log \frac{1}{\epsilon} + \sqrt{\frac{mL}{\epsilon}})\right\}^1$
Katyusha^{ns}[1]	NA	$\mathcal{O}\left\{\frac{m}{\sqrt{\epsilon}} + \sqrt{\frac{mL}{\epsilon}}\right\}$
RGEM[19]	$\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\mu}}) \log \frac{1}{\epsilon}\right\}$	NA

The *Varag* algorithm

Algorithm 1 The variance-reduced accelerated gradient (Varag) method

Input: $x^0 \in X, \{T_s\}, \{\gamma_s\}, \{\alpha_s\}, \{p_s\}, \{\theta_t\}$, and a probability distribution $Q = \{q_1, \ldots, q_m\}$ on $\{1, \ldots, m\}$.

```
Set x̃^0 = x^0.
for s = 1, 2, ... do
    Set x̃ = x̃^{s-1} and g̃ = ∇f(x̃).
    Set x_0 = x^{s-1}, x̄_0 = x̃ and T = T_s.
    for t = 1, 2, ..., T do
        Pick i_t ∈ {1, ..., m} randomly according to Q.
        x̲_t = [(1 + μγ_s)(1 − α_s − p_s)x̄_{t−1} + α_s x_{t−1} + (1 + μγ_s)p_s x̃] / [1 + μγ_s(1 − α_s)].
        G_t = (∇f_{i_t}(x̲_t) − ∇f_{i_t}(x̃))/(q_{i_t} m) + g̃.
        x_t = arg min_{x∈X} {γ_s [⟨G_t, x⟩ + h(x) + μV(x̲_t, x)] + V(x_{t−1}, x)}.
        x̄_t = (1 − α_s − p_s)x̄_{t−1} + α_s x_t + p_s x̃.
    end for
    Set x^s = x_T and x̃^s = Σ_{t=1}^T (θ_t x̄_t) / Σ_{t=1}^T θ_t.
end for
```

- Similar to SVRG algorithmic scheme
- Adopt AC-SA[GL201] in the inner loop
- Allows general distance via prox-function V
- When $\alpha_s = 1, p_s = 0$, *Varag* reduces to non-accelerated method, and achieves $\mathcal{O}\{(m + {}^{L}/_{\mu})\log({}^{1}/_{\epsilon})\}$ as SVRG.

Theorem 1 (Smooth finite-sum optimization) *Suppose that the probabilities q_i's are set to $L_i/\sum_{i=1}^m L_i$ for $i = 1,\ldots,m$, and weights $\{\theta_t\}$ are set as*

$$\theta_t = \begin{cases} \frac{\gamma_s}{\alpha_s}(\alpha_s + p_s) & 1 \le t \le T_s - 1 \\ \frac{\gamma_s}{\alpha_s} & t = T_s. \end{cases} \tag{2.2}$$

Moreover, let us denote $s_0 := \lfloor \log m \rfloor + 1$ and set parameters $\{T_s\}$, $\{\gamma_s\}$ and $\{p_s\}$ as

$$T_s = \begin{cases} 2^{s-1}, & s \le s_0 \\ T_{s_0}, & s > s_0 \end{cases}, \gamma_s = \frac{1}{3L\alpha_s}, \text{ and } p_s = \frac{1}{2}, \text{ with} \tag{2.3}$$

$$\alpha_s = \begin{cases} \frac{1}{2}, & s \le s_0 \\ \frac{2}{s - s_0 + 4}, & s > s_0 \end{cases}. \tag{2.4}$$

Then the total number of gradient evaluations of f_i performed

Katyusha$^{\text{ns}}$ [1]	$\mathcal{O}\left\{\frac{m}{\sqrt{\epsilon}} + \sqrt{\frac{mL}{\epsilon}}\right\}$

i.e., a point $\bar{x} \in X$ s.t. $\mathbb{E}[\psi(\bar{x}) - \psi^*] \le \epsilon$, can be bounded by

$$\bar{N} := \begin{cases} \mathcal{O}\left\{m \log \frac{D_0}{\epsilon}\right\}, & m \ge D_0/\epsilon, \\ \mathcal{O}\left\{m \log m + \sqrt{\frac{mD_0}{\epsilon}}\right\}, & m < D_0/\epsilon, \end{cases} \tag{2.5}$$

where D_0 is defined as

$$D_0 := 2[\psi(x^0) - \psi(x)] + 3LV(x^0, x). \tag{2.6}$$

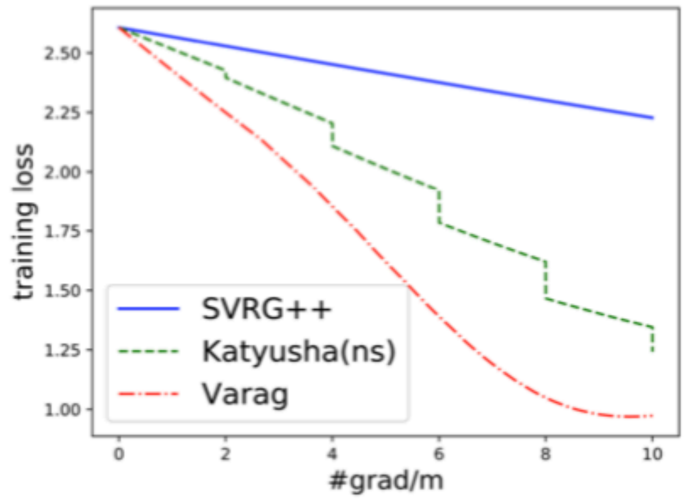
- ***Varag* solves smooth problem directly!**
  - Doubling epoch length of inner loop
  - When the required accuracy ϵ is low and/or the number of components m is large, *Varag* achieves a fast **linear rate of convergence**
  - Otherwise, *Varag* achieves an **optimal sublinear rate of convergence**

- *Varag* is the first accelerated RIG in the literature to obtain such convergence results by directly solving smooth finite-sum optimization problems.

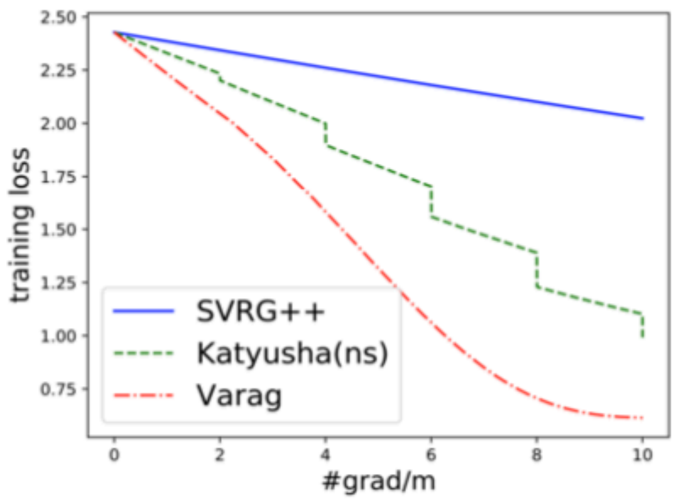
Smooth convex finite-sum optimization

One numerical example – unconstrained logistic models

$$\min_{x \in \mathbb{R}^n} \{\psi(x) := \tfrac{1}{m}\textstyle\sum_{i=1}^{m} f_i(x)\} \text{ where } f_i(x) := \log(1 + \exp(-b_i a_i^T x))\}$$



Diabetes ($m = 1151$), unconstrained logistic



Breast Cancer Wisconsin ($m = 683$), unconstrained logistic

When $\mu \approx 0$ for strongly convex problems…

When the problem is almost not strongly convex, i.e., $\mu \approx 0$, $\sqrt{mL}/_{\mu} \log(^1/_\epsilon)$ will be dominating and tend to ∞ as μ decreases.

Therefore, these complexity bounds are significantly worse than simply treating the problem as smooth convex problems.

Algorithms	Deterministic smooth strongly convex
RPDG[18]	$\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\mu}}) \log \frac{1}{\epsilon}\right\}$
Catalyst[20]	$\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\mu}}) \log \frac{1}{\epsilon}\right\}^1$
Katyusha[1]	$\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\mu}}) \log \frac{1}{\epsilon}\right\}$
Katyusha^{ns}[1]	NA
RGEM[19]	$\mathcal{O}\left\{(m + \sqrt{\frac{mL}{\mu}}) \log \frac{1}{\epsilon}\right\}$

$$\bar{N} := \begin{cases} \mathcal{O}\left\{m \log \frac{D_0}{\epsilon}\right\}, & m \geq D_0/\epsilon, \\ \mathcal{O}\left\{m \log m + \sqrt{\frac{mD_0}{\epsilon}}\right\}, & m < D_0/\epsilon, \end{cases}$$

Theorem 2 (A unified result for convex finite-sum optimization) *Suppose that the probabilities q_i's are set to $L_i/\sum_{i=1}^m L_i$ for $i = 1, \ldots, m$. Moreover, let us denote $s_0 := \lfloor \log m \rfloor + 1$ and assume that the weights $\{\theta_t\}$ are set to (2.2) if $1 \le s \le s_0$ or $s_0 < s \le s_0 + \sqrt{\frac{12L}{m\mu}} - 4$, $m < \frac{3L}{4\mu}$. Otherwise, they are set to*

$$\theta_t = \begin{cases} \Gamma_{t-1} - (1 - \alpha_s - p_s)\Gamma_t, & 1 \le t \le T_s - 1, \\ \Gamma_{t-1}, & t = T_s, \end{cases} \tag{2.7}$$

where $\Gamma_t = (1 + \mu\gamma_s)^t$. If the parameters $\{T_s\}$, $\{\gamma_s\}$ and $\{p_s\}$ set to (2.3) with

$$\boxed{\alpha_s = \begin{cases} \frac{1}{2}, & s \le s_0, \\ \max\left\{\frac{2}{s - s_0 + 4}, \min\{\sqrt{\frac{m\mu}{3L}}, \frac{1}{2}\}\right\}, & s > s_0, \end{cases}} \tag{2.8}$$

then the total number of gradient evaluations of f_i performed by Algorithm 1 to find a stochastic ϵ-solution of (1.1) can be bounded by

$$\boxed{\bar{N} := \begin{cases} \mathcal{O}\left\{m \log \frac{D_0}{\epsilon}\right\}, & m \ge \frac{D_0}{\epsilon} \text{ or } m \ge \frac{3L}{4\mu}, \\ \mathcal{O}\left\{m \log m + \sqrt{\frac{mD_0}{\epsilon}}\right\}, & m < \frac{D_0}{\epsilon} \le \frac{3L}{4\mu}, \\ \mathcal{O}\left\{m \log m + \sqrt{\frac{mL}{\mu}} \log \frac{D_0/\epsilon}{3L/4\mu}\right\}, & m < \frac{3L}{4\mu} \le \frac{D_0}{\epsilon}. \end{cases}} \tag{2.9}$$

wher

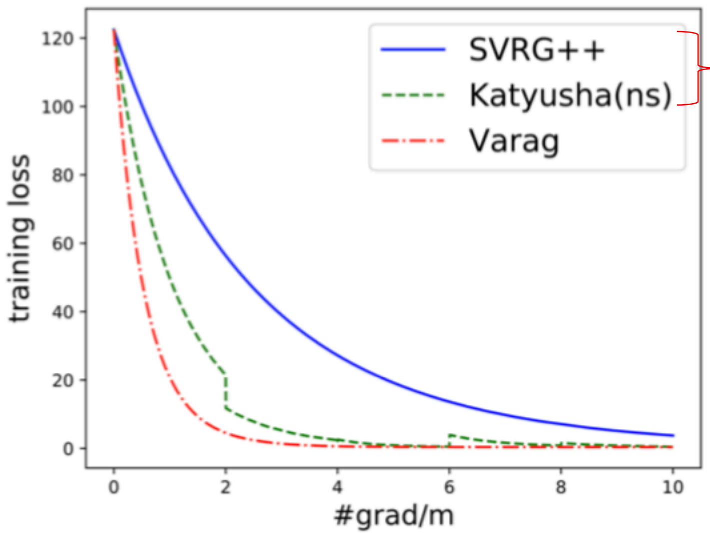
$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \mu V(x, y), \forall x, y \in X.$$

- ***Varag* is an unified optimal method!**
  - When μ is large enough, *Varag* achieves the optimal linear rate of convergence
  - When μ is relatively small, *Varag* treats the problem as a smooth convex problem.
- *Varag* does not require to know the target accuracy ϵ and the constant D_0 beforehand to obtain optimal convergent rates.
- The unified step-size policy adjusts itself to the value of the condition number
- *Varag* does not assume the strong convexity comes from the regularizer!

Varag as an unified optimal method

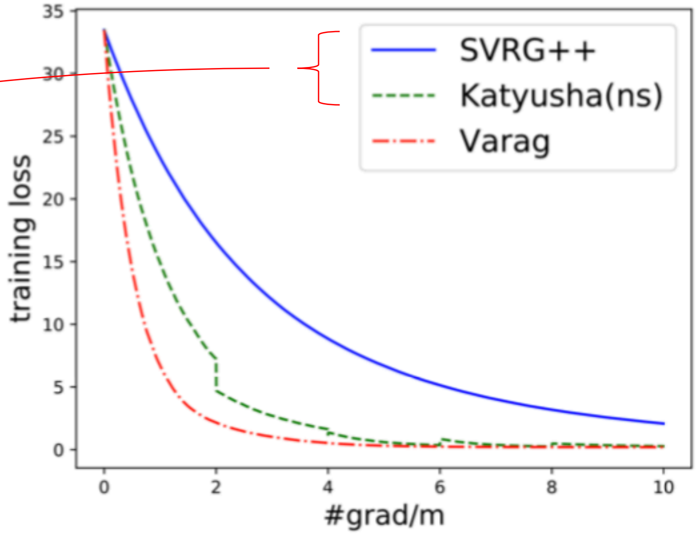
One numerical example – Lasso regression models

$$\min_{x\in\mathbb{R}^n}\{\psi(x) := \frac{1}{m}\sum_{i=1}^m f_i(x) + h(x)\} \text{ where } f_i(x) := \frac{1}{2}(a_i^T x - b_i)^2, h(x) := \lambda\|x\|_1.$$



Diabetes ($m = 1151$), Lasso $\lambda = 0.001$

Treats it as smooth convex problems



Breast Cancer Wisconsin ($m = 683$), Lasso $\lambda = 0.001$

$$V(x, X^*) \le \frac{1}{\bar{\mu}}(\psi(x) - \psi^*),\ \forall x \in X,$$

Theorem 3 (Convex finite-sum optimization under error bound) *Ass*
q_i's are set to $L_i / \sum_{i=1}^m L_i$ for $i = 1, \ldots, m$, and θ_t are defined as (
parameters $\{\gamma_s\}$, $\{p_s\}$ and $\{\alpha_s\}$ as in (2.3) and (2.4), $s = 4 + 4\sqrt{\frac{L}{\bar{\mu} m}}$,

Moreover, if we restart Varag *every time it runs s iterations for $k = \log$*
number of gradient evaluations of f_i to find a stochastic ϵ-solution of (1.

$$\bar{N} := k(\textstyle\sum_s (m + T_s)) = \mathcal{O}\left\{\left(m + \sqrt{\frac{mL}{\bar{\mu}}}\right) \log \frac{\psi(x^0) - \psi(x^*)}{\epsilon}\right\}. \tag{2.13}$$

Application examples:
Linear systems, quadratic programs, linear matrix inequalities and composite problems, etc.

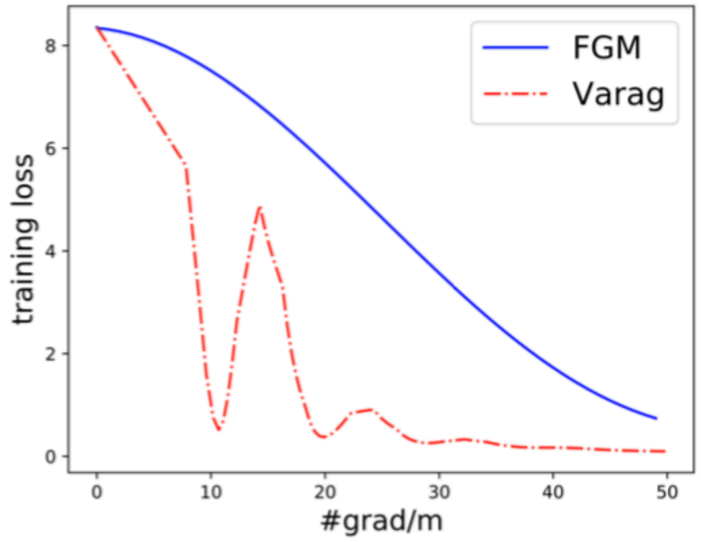
Varag is **the first randomized** method to establish the **accelerated linear rate of convergence** for solving the above problems!

Generalization of *Varag*

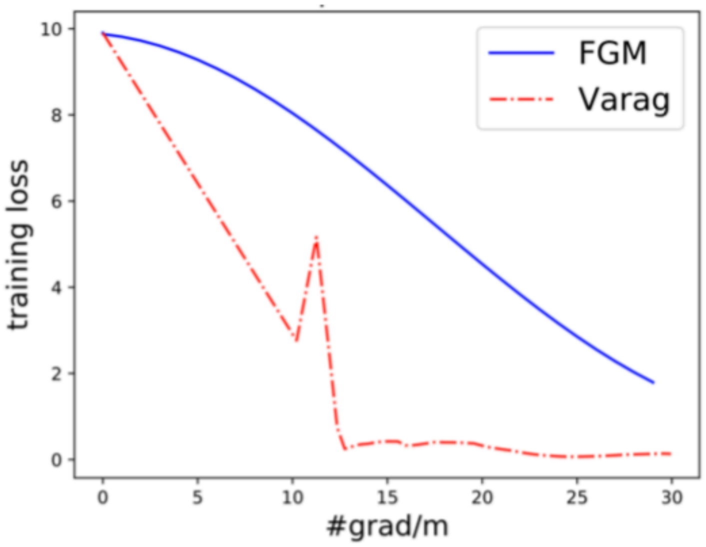
Finite-sum under error bound condition

One numerical example – quadratic problems

$$\min_{x \in \mathbb{R}^n} \{\psi(x) := \tfrac{1}{m}\textstyle\sum_{i=1}^{m} f_i(x)\} \text{ where } f_i(x) := \tfrac{1}{2} x^T Q_i x + q_i^T x.$$



Diabetes ($m = 1151$)



Parkinsons Telemonitoring ($m = 5875$)

Only noisy gradient information can be accessed via SFO

$$\mathbb{E}_{\xi_j}[G_i(x,\xi_j)] = \nabla f_i(x),\ i = 1,\dots,m,$$
$$\mathbb{E}_{\xi_j}[\|G_i(x,\xi_j) - \nabla f_i(x)\|_*^2] \le \sigma^2,\ i = 1,\dots,m.$$

Algorithm 2 Stochastic accelerated variance-reduced stochastic gradient descent (Stochastic Varag)

This algorithm is the same as Algorithm 1 except that for given batch-size parameters B_s and b_s, Line 3 is replaced by $\tilde{x} = \tilde{x}^{s-1}$ and

$$\tilde{g} = \tfrac{1}{m}\textstyle\sum_{i=1}^{m}\left\{G_i(\tilde{x}) := \tfrac{1}{B_s}\textstyle\sum_{j=1}^{B_s} G_i(\tilde{x},\xi_j^s)\right\}, \tag{2.16}$$

and Line 8 is replaced by

$$G_t = \tfrac{1}{q_{i_t} m b_s}\textstyle\sum_{k=1}^{b_s}\left(G_{i_t}(\underline{x}_t,\xi_k^s) - G_{i_t}(\tilde{x})\right) + \tilde{g}. \tag{2.17}$$

Generalization of *Varag*

Stochastic finite-sum

Theorem 4 (Stochastic smooth finite-sum optimization) *Assume that θ_t are defined as in (2.2), $C := \sum_{i=1}^{m} \frac{1}{q_i m^2}$ and the probabilities q_i's are set to $L_i / \sum_{i=1}^{m} L_i$ for $i = 1, \ldots, m$. Moreover, let us denote $s_0 := \lfloor \log m \rfloor + 1$ and set T_s, α_s, γ_s and p_s as in (2.3) and (2.4). Then the number of calls to the SFO oracle required by Algorithm 2 to find a stochastic ϵ-solution of (1.1) can be bounded by*

$$\boxed{N_{\text{SFO}} = \sum_s (mB_s + T_s b_s) = \begin{cases} \mathcal{O}\left\{\frac{mC\sigma^2}{L\epsilon}\right\}, & m \geq D_0/\epsilon, \\ \mathcal{O}\left\{\frac{C\sigma^2 D_0}{L\epsilon^2}\right\}, & m < D_0/\epsilon, \end{cases}} \tag{2.18}$$

where D_0 is given in (2.6).

Varag is **the first** to achieve the above complexity results for smooth convex problems!

- RGEM[LZ18] achieves nearly optimal rate $\tilde{\mathcal{O}}\{{}^{\sigma^2}/_{\mu^2\epsilon}\}$ for expected distance between the output and the optimal solution
- Variant of SVRG[KM19] achieves $\mathcal{O}\{m\text{log}m + {}^{\sigma\text{\^{}}2}/_{\epsilon}\}$ with some specific initial point.

Future works

- Extend *Varag* to solve nonconvex finite-sum problems
- How to choose stepsize if L and μ are hard to estimate?

Thank you!

References

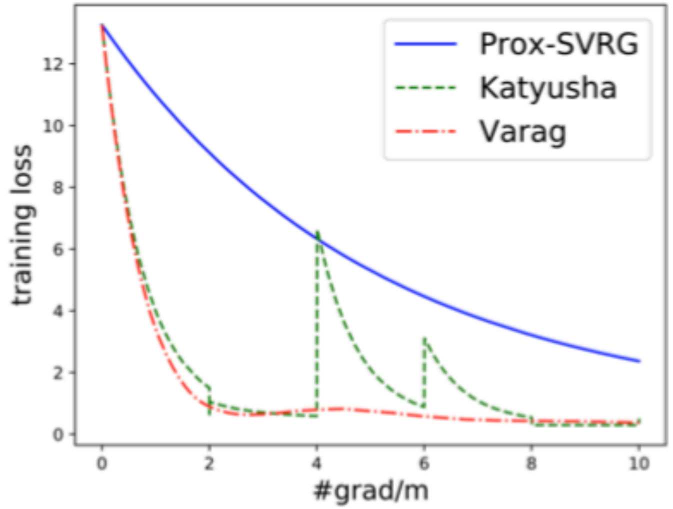
Varag*:* Lan, G., Li, Z., & Zhou, Y. (2019). *A unified variance-reduced accelerated gradient method for convex optimization*. *arXiv preprint arXiv:1905.12412*. Accepted by NeurIPS 2019.

Other RIG methods:

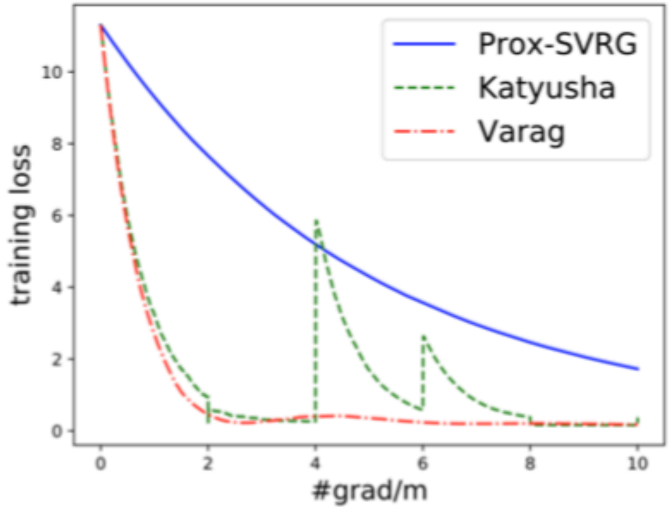
- Johnson, R., & Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems* (pp. 315-323).
- Xiao, L., & Zhang, T. (2014). A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, *24*(4), 2057-2075.
- Defazio, A., Bach, F., & Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 1646-1654.
- Nguyen, L. M., Liu, J., Scheinberg, K., & Takáč, M. (2017, August). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 2613-2621). JMLR. org.
- Allen-Zhu, Z., & Yuan, Y. (2016, June). Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In *International conference on machine learning* (pp. 1080-1089).
- Lan, G., & Zhou, Y. (2018). An optimal randomized incremental gradient method. *Mathematical programming*, *171*(1-2), 167-215.
- Allen-Zhu, Z. (2017). Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, *18*(1), 8194-8244.
- Lin, H., Mairal, J., & Harchaoui, Z. (2015). A universal catalyst for first-order optimization. *Advances in neural information processing systems*, 3384-3392.
- Lan, G., & Zhou, Y. (2018). Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, *28*(4), 2753-2782.
- Kulunchakov, A., & Mairal, J. (2019). Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *arXiv preprint arXiv:1901.08788*.

One numerical example – ridge regression models

$$\min_{x \in \mathbb{R}^n} \{\psi(x) := \tfrac{1}{m}\textstyle\sum_{i=1}^m f_i(x) + h(x)\} \text{ where } f_i(x) := \tfrac{1}{2}(a_i^T x - b_i)^2, h(x) := \lambda\|x\|_2^2.$$



Diabetes ($m = 1151$), ridge $\lambda = 10^{-6}$



Breast-Cancer-Wisconsin ($m = 683$), ridge $\lambda = 10^{-6}$

Varag requires less CPU time per training epoch than Katyusha!