Introductory Lectures on Optimization

Stochastic Optimization (2)

Hui Qian
qianhui@zju.edu.cn

College of Computer Science, Zhejiang University

December 16, 2024

Outline

Part III
Principles for Improving SGD

Improving SGD

Although it is very hard to improve the convergence rate of the Stochastic Gradient Descent (SGD) method, we can still try to accelerate the performance by improving the constant factors. We discuss several strategies below.

Reduce Variance

1. Mini-Batch Sampling: use a small batch of samples instead of one to estimate the gradient at every iteration

$$G(x_t, \xi_t) \Rightarrow \frac{1}{b}\sum_{i=1}^{b} G(x_t, \xi_{t,i}).$$

Consequently, the variance of the new stochastic gradient will be $\mathcal{O}(b)$ times smaller, i.e. the constant term M^2 in the convergence now reduces to M^2/b.

Improving SGD

Remark.

Result 1 (for fixed step-size):

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{\gamma L \delta_g^2}{2\mu} + (1-\gamma\mu)^t (f(x_1) - f(x^*)).$$

where

$$\mathbb{E}\left[\|g(x,\xi)^2\|_2\right] \leq \delta_g^2 + c_g \left\|\nabla f(x)\right\|_2^2.$$

Result 2 (for shrinking step-size):

$$\mathbb{E}\left[\|x_t - x_*\|_2^2\right] \leq \frac{C(\gamma)}{t}$$

where

$$C(\gamma) = \max\left\{\frac{\gamma^2 M^2}{2\mu\gamma - 1}, \|x_1 - x_*\|_2^2\right\}.$$

Improving SGD

Reduce Variance (continued...)

2. Importance Sampling: Instead of sampling form $\xi \sim \mathbb{P}$, we can obtain samples from another well defined random variable η with nominal distribution Q, and use a different stochastic gradient,

$$G(x_t, \xi_t) \Rightarrow G(x_t, \eta_t)\frac{P(\eta_t)}{Q(\eta_t)}.$$

The variance of the new stochastic gradient under properly chosen distribution Q could be smaller[Zhao and Zhang, 2015; Needell et al., 2014].

3. Stratified Sampling: Another line of research that also aims to effectively reduce the variance of gradient estimates is stratified sampling [Zhao and Zhang, 2014].
4. Using historical information: SVRG, SARAH, SPIDER, STORM, etc.

Improving SGD

Remark.

Suppose we wish to estimate $g = \mathbb{E}_{\xi\sim P}[h(\xi)]$, where $h(\xi) = G(x,\xi)$. Let Q be another PDF with the proerty that $Q(\xi) \neq 0$ whenever $P(\xi) \neq 0$. Then

$$g = \mathbb{E}_P[h(\xi)] = \int h(\xi)P(\xi)d\xi = \int h(\xi)\frac{P(\xi)}{Q(\xi)}Q(\xi)d\xi = \mathbb{E}_Q\left[\underbrace{h(\xi)\frac{P(\xi)}{Q(\xi)}}_{h^*(\xi)}\right].$$

Also, we have

$$\mathrm{Var}_P(h(\xi)) - \mathrm{Var}_Q(h^*(\xi)) = \int h(\xi)^2 P(\xi)\left(1 - \frac{P(\xi)}{Q(\xi)}\right)d\xi.$$

Improving SGD

Adaptive Stepsize

The traditional fixed stepsize $\gamma_t = 1/\mu t$ may be too small so that the efficiency of the stochastic gradient descent approach can be compromised. One may instead select the stepsize adaptively to optimize the progress at each iteration. For instance, in [Yousefian et al., 2012], the authors propose to automatically update the stepsize based on the recursion

$$\gamma_t = \frac{1}{\mu t} \Rightarrow \gamma_t = \gamma_{t-1}(1 - c\gamma_{t-1}).$$

Using Bregman Distance

Stochastic mirror descent (SMD) and its variants make up arguably one of the most widely used families of first-order methods in stochastic optimization.

Part IV
Stochastic Mirror Descent

Stochastic Mirror Descent

Analogous to the deterministic optimization scenario, Mirror Descent Stochastic Approximation (a.k.a. Stochastic Mirror Descent) is adopted to solve non-smooth problems.

Let $w(x)$ be a continuously differentiable and 1-strongly convex function w.r.t. some norm $\|\cdot\|$. A simple example of a distance-generating function is $w(x) = \frac{1}{2}\|x\|_2^2$. Define function $V(x, y) = w(x) - w(y) - \nabla w(y)^\top (x - y)$, which is called the Bregman distance.

The mirror descent stochastic approximation works as follows:

$$x_{t+1} = \arg\min_{x \in X} \{V(x, x_t) + \langle \gamma_t G(x_t, \xi_t),\ x \rangle\}.$$

Stochastic Mirror Descent

Theorem 1 (Nemirovski et al. [2009])

Let f be a convex function, $\Omega = \max_{x \in X} V(x, x_1)$. Let the candidate solution $\hat{x}_T$ be the weighted average

$$\hat{x}_T = \sum_{t=1}^{T} \gamma_t x_t \Big/ \sum_{t=1}^{T} \gamma_t$$

If there exists $M > 0$, s.t., $\mathbb{E}[\|G(x,\xi)\|_*^2] \leq M^2, \forall x \in X$, then

$$\mathbb{E}[(f(\hat{x}_T) - f(x_*)] \leq \frac{\Omega + \frac{M^2}{2} \sum_{t=1}^{T} \gamma_t^2}{\sum_{t=1}^{T} \gamma_t}.$$

Stochastic Mirror Descent

$$\boxed{\mathbb{E}[(f(\hat{x}_T) - f(x_*)] \leq \frac{\Omega + \frac{M^2}{2}\sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t}.}$$

Proof. Based on the optimality condition of the mirror descent stochastic approximation (see (2.36) of [Nemirovski et al., 2009]), we have

$$\gamma_t(x_t - x_*)^\top G(x_t, \xi_t) \leq V(x_t, x_*) - V(x_{t+1}, x_*) + \frac{\gamma_t^2}{2}\|G(x_t, \xi_t)\|_*^2$$

Rewrite the above as follows (where $g(x_t) \in \partial f(x_t)$).

$$\begin{aligned}\gamma_t(x_t - x_*)^\top g(x_t) \leq\ & V(x_t, x_*) - V(x_{t+1}, x_*) \\ & - \gamma_t(G(x_t, \xi_t) - g(x_t))^\top (x_t - x_*) + \frac{\gamma_t}{2}\|G(x_t, \xi_t)\|_*^2.\end{aligned}$$

Stochastic Mirror Descent

Proof. (continued) Taking summation over $t = 1, \dots, T$, we have

$$\sum_{t=1}^{T} \gamma_t (x_t - x_*)^\top g(x_t) \le V(x_1, x_*) + \sum_{t=1}^{T} \frac{\gamma_t^2}{2} \|G(x_t, \xi_t)\|_*^2 - \sum_{t=1}^{T} \gamma_t (G(x_t, \xi_t) - g(x_t))^\top (x_t - x_*). \quad (1)$$

Let's set $\hat{x}_T = \frac{\sum_{t=1}^{T} \gamma_t x_t}{\sum_{t=1}^{T} \gamma_t}$, and onsider the convexity of $f(x)$, we have

$$\sum_{t=1}^{T} \gamma_t (x_t - x_*)^\top g(x_t) \ge \sum_{t=1}^{T} \gamma_t (f(x_t) - f(x_*)) \ge \left(\sum_{t=1}^{T} \gamma_t \right) (f(\hat{x}_T) - f(x_*)). \quad (2)$$

Stochastic Mirror Descent

Proof. (continued) Combine (1) and (2), we can get

$$\begin{aligned} f(\hat{x}_T) - f(x_*) \leq & \frac{V(x_1, x_*) + \sum_{t=1}^{T} \frac{\gamma_t^2}{2} \|G(x_t, \xi_t)\|_*^2}{\sum_{t=1}^{T} \gamma_t} \\ & - \frac{\sum_{t=1}^{T} \gamma_t (G(x_t, \xi_t) - g(x_t))^\top (x_t - x_*)}{\sum_{t=1}^{T} \gamma_t}. \end{aligned} \tag{3}$$

Taking expectations on both sides of (3), we can have

$$\mathbb{E}[f(\hat{x}_T) - f(x_*)] \leq \frac{\max_{x \in X} V(x_1, x) + \frac{M^2}{2} \sum_{t=1}^{T} \gamma_t^2}{\sum_{t=1}^{T} \gamma_t}$$

as desired. $\square$

Improving SMD: Adaptive Bregman Distance

One may also adaptively choose the Bregman distance and hope to improve the efficiency. For instance, the AdaGrad algorithm in [Duchi et al., 2011] propose the following

$$w(x) = x^\top x \Rightarrow w_t(x) = \frac{1}{2} x^\top H_t x,$$

where $H_t = \delta \boldsymbol{I} + [\sum_{k=1}^t g_k g_k^\top]^{\frac{1}{2}}$, and $g_t = G(x_t, \xi_t)$.

Remark: some variants including Adadelta[Zeiler, 2012], RMSProp[Hinton et al., 2012]. Other popular first order solvers are Adam [Kingma and Ba, 2014], Adaptive Moment Estimation, (and its variant Nadam [Dozat, 2016], Nesterov-accelerated Adaptive Moment Estimation), which also reduce the radically diminishing learning rates of Adagrad.

Part V
Variance Reduction Methods

Finite Sum Problems Revisit

Problems where the objective function can be defined as a finite sum of functions, are called finite sum problems, or big-n problem. Formally, a finite sum problem can be written as,

$$\min_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \psi(x).$$

Notice that the structure is similar to sample average approximation (SAA) described in the lectures above.

The number of functions, n, is analogous to the sample size drawn in Monte Carlo sampling ($i.i.d.$ samples).

Finite Sum Problems Revisit

Such type of problems are popular in many applications.

1. Empirical risk minimizatin: In machine learning problems, the risk associated with a hypothesis (h) is approximated by an empirical risk $R(h)$, defined as the loss over the dataset $(x_1, y_1), \ldots, (x_n, y_n)$. Empirical risk given by

$$R(h_\theta) = \frac{1}{n}\sum_{i=1}^{n} L(h_\theta(x_i), y_i)),$$

has the structure of a finite sum problem.

2. Distributed optimization: Distributed optimization involves a finite sum problem being solved by a group of computational entities (agents).By using an iterative consensus and local gradient based algorithm, one can show the convergence of local state estimate to the optimum.

Variance Reduction Techniques

Suppose we want to estimate $\Theta = \mathbb{E}[X]$, the expected value of a random variable X. Suppose we also have access to a random variable Y which is highly correlated with X, and we can compute $\mathbb{E}[Y]$ easily. Let's consider the following point estimator $\hat{\Theta}_\alpha$ with $\alpha \in [0, 1]$:

$$\hat{\Theta}_\alpha = \alpha(X - Y) + \mathbb{E}[Y].$$

The expectation and variance are given by,

$$\begin{aligned}
\mathbb{E}[\hat{\Theta}_\alpha] &= \alpha\mathbb{E}[X] + (1 - \alpha)\mathbb{E}[Y] \\
\mathrm{Var}[\hat{\Theta}_\alpha] &= \alpha^2(\mathrm{Var}[X] + \mathrm{Var}[Y] - 2\mathrm{Cov}[X, Y]).
\end{aligned}$$

Variance Reduction Techniques

Note that, choosing a suitable α, we can achieve a balance between variance and bias.

1. $\alpha = 1$, this estimator becomes $(X - Y) + \mathbb{E}[Y]$, which is an unbiased estimator.
2. $\alpha = 0$, this estimator reduces to a constant $\mathbb{E}[Y]$, which has zero variance but could be heavily biased.
3. if $\text{Cov}[X, Y]$ is sufficiently large, then $\text{Var}[\hat{\Theta}_\alpha] < \text{Var}[X]$. The new estimator $\hat{\Theta}_\alpha$ has smaller variance than the direct estimator X.
4. As α increases from 0 to 1, the bias decreases and the variance increases.

Recently develeped incremental gradient algorithms namely SAG, SAGA, SVRG and S2GD are all special cases of the general variance reduction technique described above.

SVRG

For convex and L-smooth function f_i, Consider

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x),$$

where f is μ-strongly.
Main idea of SVRG: The stochastic variance reduced gradient method (SVRG)[Johnson and Zhang, 2013], the prototypical snapshot method, using a full-gradient that is reevaluated at a snapshot point at regular intervals of m iteration. Specifically, if we have access to a history point x^{old} and $\nabla F(x^{\text{old}})$, then the gradient estimation at x^t is

$$\underbrace{\nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{\text{old}})}_{\to 0 \text{ if } x^t \approx x^{\text{old}}} + \underbrace{\nabla f(x^{\text{old}})}_{\to 0 \text{ if } x^{\text{old}} \approx x^*}$$

with $i_t \sim \text{Unif}(1, \ldots, n)$.

SVRG

Intuition of SVRG: If the current iterate is not too far away from previous iterates, then historical gradient info might be useful in producing a better estimator to reduce variance.

1. SVRG does not require storage of gradient as seen in SAG or SAGA.
2. Convergence rates for SVRG can be proved easily and a very intuitive explanation can be provided by linking increased speed to reduced variance.

Remark.
(1) unbiased estimate of ∇f; (2) variability is reduced.

SVRG

Parameters: update frequency m and learning rate η
Initialize $\tilde{x}_0$
for $s = 1, 2, \ldots$ **do**
  $\tilde{x} = \tilde{x}^{s-1}$
  $\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x})$
  $x_0 = \tilde{x}$
  for $t = 1, 2, \ldots, m$ **do**
    Randomly pick $i_t \in \{1, 2, \ldots, n\}$ and update weight,
    $$x^t = x^{t-1} - \eta\left(\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(\tilde{x}) + \tilde{\theta}\right)$$
  end for
  Update $\tilde{x}^s = \frac{1}{m}\sum_{t=1}^{m} x^t$
end for

Theoretical Analysis of SVRG

Lemma 2

Assume $f_i(x)$ is convex and L-smooth. For any x, we have

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x) - \nabla f_i(x_*)\|_2^2 \leq 2L(f(x) - f(x_*)). \tag{4}$$

Proof. For any i, consider a function $g_i(x)$,

$$g_i(x) = f_i(x) - f_i(x_*) - \nabla f_i(x_*)^\top (x - x_*).$$

Note that $\nabla g_i(x) = \nabla f_i(x) - \nabla f_i(x_*)$. Clearly, $\nabla g_i(x_*) = 0$, implying that $g_i(x_*) = \min_x g_i(x)$.

SVRG

Proof. (continued) Therefore, following the definition of minimum and the L-smoothness of function $g_i(x)$, we arrive at

$$0 = g_i(x_*) \leq \min_{\eta}[g_i(x - \eta \nabla g_i(x))] \leq g_i(x) - \frac{1}{2L} \|\nabla g_i(x)\|_2^2 .$$

That is,

$$\|\nabla g_i(x)\|_2^2 \leq 2L g_i(x) \quad \text{(the following is by def. of } g_i(x))$$
$$\|\nabla f_i(x) - \nabla f_i(x_*)\|_2^2 \leq 2L(f_i(x) - f_i(x_*) - \nabla f_i(x_*)^\top (x - x_*)).$$

By summing the above inequality and using the fact $\nabla f(x_*) = 0$ and $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(x_*)\|_2^2 \leq 2L(f(x) - f(x_*)).$$

Theoretical Analysis of SVRG

Theorem 3

Assume $f_i(x)$ is convex and L-smooth and $f(x)$ is μ-strongly convex. Let $x_* = \text{argmin}\, f(x)$. Assume m is sufficiently large (and $\eta < \frac{1}{2L}$), so that

$$\rho = \frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1,$$

then we have geometric convergence in expectation for SVRG, i.e.,

$$\mathbb{E}[f(\tilde{x}^s) - f_*] \leq \rho^s [f(\tilde{x}^0) - f_*].$$

Theoretical Analysis of SVRG

Proof. Let $v^t = \nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(\tilde{x}) + \nabla f(\tilde{x})$. We now take expectation with respect to i_t conditioned on x^{t-1} and obtain,

$$\mathbb{E}[\left\|v^t\right\|^2] = \mathbb{E}[\left\|[\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x_*)] + [\nabla f_{i_t}(x_*) - \nabla f_{i_t}(\tilde{x}) + \nabla f(\tilde{x})]\right\|_2^2]$$

$$(\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2)$$

$$\le 2\mathbb{E}[\left\|\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x_*)\right\|_2^2] + 2\mathbb{E}[\|\nabla f_{i_t}(\tilde{x}) - \nabla f_{i_t}(x_*) - \nabla f(\tilde{x})\|_2^2]$$

$$(\nabla f(\tilde{x}) = \mathbb{E}[\nabla f_{i_t}(\tilde{x}) - \nabla f_{i_t}(x_*)]) \text{ and } (\mathbb{E}[\|e - \mathbb{E}[e]\|_2^2] \le \mathbb{E}[\|e\|_2^2])$$

$$\le 2\mathbb{E}[\left\|\nabla f_{i_t}(x^{t-1}) - \nabla f_{i_t}(x_*)\right\|_2^2] + 2\mathbb{E}[\|\nabla f_{i_t}(\tilde{x}) - \nabla f_{i_t}(x_*)\|_2^2]$$

$$\le 4L[f(x^{t-1}) - f(x_*) + f(\tilde{x}) - f(x_*)] \quad (\text{ by } (4)) \tag{5}$$

Theoretical Analysis of SVRG

Proof. (continued) Now notice from the definition of v^t, $\mathbb{E}[v^t|x^{t-1}] = \nabla f(x^{t-1})$; and this leads to,

$$
\begin{aligned}
\mathbb{E}[\|x^t - x_*\|_2^2] &= \mathbb{E}[\|\boxed{x^{t-1} - \eta v^t} - x_*\|_2^2](\text{ one step descent}) \\
&= \|x^{t-1} - x_*\|_2^2 - 2\eta(x^{t-1} - x_*)^\top \mathbb{E}[v^t] + \eta^2 \mathbb{E}[\|v^t\|_2^2](\text{ conditioned on } x^{t-1}) \\
&\leq \|x^{t-1} - x_*\|_2^2 - 2\eta(x^{t-1} - x_*)^\top \nabla f(x^{t-1}) \\
&\quad + 4L\eta^2[f(x^{t-1}) - f(x_*) + f(\tilde{x}) - f(x_*)](\text{by } (5)) \\
&\leq \|x^{t-1} - x_*\|_2^2 - 2\eta(f(x^{t-1}) - f(x_*)) + 4L\eta^2[f(x^{t-1}) - f(x_*) + f(\tilde{x}) - f(x_*)] \\
&= \|x^{t-1} - x_*\|_2^2 - 2\eta(1 - 2L\eta)(f(x^{t-1}) - f(x_*)) + 4L\eta^2[f(\tilde{x}) - f(x_*)].
\end{aligned}
$$

We consider a fixed stage s, so that $\tilde{x} = \tilde{x}^{s-1}$ and $\tilde{x}^s$ is selected after all the updates have completed.

Theoretical Analysis of SVRG

Proof. (continued) By summing the previous inequality, taking expectation with all the history, we obtain

$$
\begin{aligned}
\mathbb{E}[\|x^m - x_*\|^2] + 2\eta(1-2L\eta)m\mathbb{E}[f(\tilde{x}^s) - f(x_*)] \\
&\le \mathbb{E}[\|x^0 - x_*\|^2] + 4Lm\eta^2\mathbb{E}[f(\tilde{x}^{s-1}) - f(x_*)] \\
&= \mathbb{E}[\|\tilde{x}^{s-1} - x_*\|^2] + 4Lm\eta^2\mathbb{E}[f(\tilde{x}^{s-1}) - f(x_*)] \\
&\le \frac{2}{\mu}\mathbb{E}[f(\tilde{x}^{s-1}) - f(x_*)] + 4Lm\eta^2\mathbb{E}[f(\tilde{x}^{s-1}) - f(x_*)]
\end{aligned}
$$

Remark: (1)$mf(\tilde{x}^s) = mf(\frac{1}{m}\sum_{i=1}^m x^{t-1}) \le \sum_{i=1}^m f(x^{t-1})$ (2)The last inequality is due to the fact that $f(x)$ is μ-strongly convex.

Theoretical Analysis of SVRG

Proof. (continued) We now have:

$$2\eta(1-2L\eta)m\mathbb{E}[f(\tilde{x}^s)-f(x_*)] \leq \left(\frac{2}{\mu}+4Lm\eta^2\right)\mathbb{E}[f(\tilde{x}^{s-1})-f(x_*)].$$

Clearly, from the above inequality we get

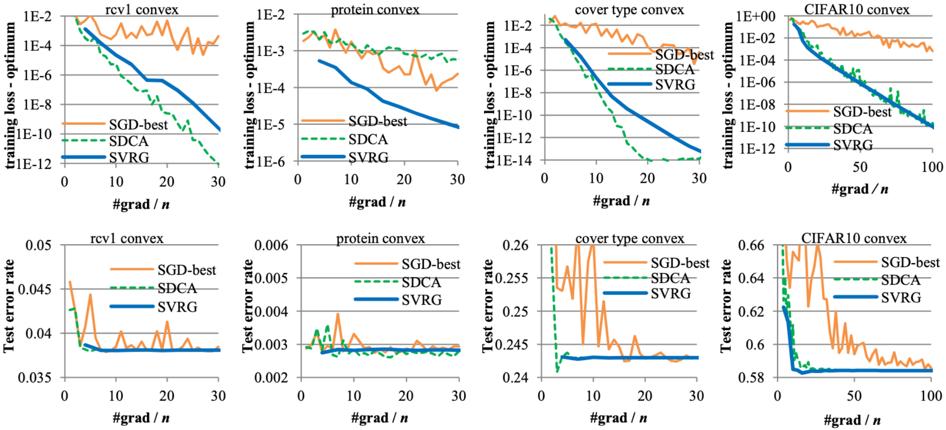
$$\mathbb{E}[f(\tilde{x}^s)-f(x_*)] \leq \left[\frac{1}{\mu\eta(1-2L\eta)m}+\frac{2L\eta}{1-2L\eta}\right]\mathbb{E}[f(\tilde{x}^{s-1})-f(x_*)]$$

This give us the desired gemoetric convergence rate,

$$\mathbb{E}[f(\tilde{x}^s)-f_*] \leq \rho^s[f(\tilde{x}^0)-f_*].$$

$\square$

Empirical Study of SVRG



ℓ_2-regularized logistic regression on CIFAR-10

Extension of SVRG

1. Non-uniform sampling: SVRG algorithm assumes uniform sampling, however, one may choose an adaptive sampling rate,

$$\mathbb{P}(i_t = i) = \frac{L_i}{\sum L_i}$$

where L_i is the smoothness parameter for f_i. This sampling strategy improves the complexity from $O((n + \frac{L_{max}}{\mu}) \log(\frac{1}{\epsilon}))$ to $O((n + \frac{L_{avg}}{\mu}) \log(\frac{1}{\epsilon}))$. Intuitively, the function $f_i(x)$ that has a higher Lipschitz constant (which is prone to change relatively rapidly) gets higher probability of getting selected.

Extension of SVRG

2. Composite convex minimization: These are problems of the form

$$\min_x \frac{1}{n}\sum_i f_i(x) + g(x)$$

where $f_i(x)$ are smooth and convex, but $g(x)$ is convex but possibly nonsmooth. Such problems can be handle by prox-SVRG [Xiao and Zhang, 2014] by imposing an additional proximal operator of g at iteration.

3. Acceleration: We can accelerate SVRG further to arrive at an optimal complexity of

$$O\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{1}{\epsilon}\right)\right)$$

This improvement is significant in problems where $\frac{L}{\mu} >> n$.

SARAH

For f_i is a L-smooth function (potentially nonconvex) for $i = 1, \ldots, n$, consider

$$\min_{x \in \mathbb{R}^d} F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

Key idea of SARAH [Nguyen et al., 2017, 2019]: recursive / adaptive update of gradent estimates

$$\begin{aligned} g^0 &= \nabla F(x^0). \\ g^t &= \nabla f_{i_t}(x^t) - \nabla f_{i_t}(x^{t-1}) + g^{t-1}. \\ x^{t+1} &= x^t - \eta g^t. \end{aligned}$$

SARAH

(1) Biased estimate of $\nabla F(x^t)$:

We have

$$\mathbb{E}[g^t|\text{everything prior to } x^t] = \nabla F(x^t) - \nabla F(x^{t-1}) + g^{t-1}$$

where $\nabla F(x^{t-1}) - g^{t-1} \neq 0$.
However if we average out all randomness, we have

$$\mathbb{E}[g^t] = \mathbb{E}[\nabla F(x^t)].$$

Remark: see [Nguyen et al., 2017] for proof.

SARAH

(2) Need reset:

For many (e.g. strongly cnovex) problems, recursive gradient estimate g^t may decay fast (variance $\downarrow$ and bias $\uparrow$).

1. g^t may quickly deviate form the target gradeint $\nabla F(x^t)$.
2. progress stalls as g^t cannot guarantee sufficient descent.

solution: reset g^t every few iterations to calibrate with the true batch gradient

SARAH

Algorithm 12.4 SARAH (Nguyen et al. '17)

1: **for** $s = 1, 2, \ldots, S$ **do**

2: $\quad \boldsymbol{x}_s^0 \leftarrow \boldsymbol{x}_{s-1}^{m+1}$, and compute $\underbrace{\boldsymbol{g}_s^0 = \nabla F(\boldsymbol{x}_s^0)}_{\text{batch gradient}}$ // restart $\boldsymbol{g}$ anew

3: $\quad \boldsymbol{x}_s^1 = \boldsymbol{x}_s^0 - \eta \boldsymbol{g}_s^0$

4: $\quad$ **for** $t = 1, \ldots, m$ **do**

5: $\quad\quad$ choose i_t uniformly from $\{1, \ldots, n\}$

6: $\quad\quad \boldsymbol{g}_s^t = \underbrace{\nabla f_{i_t}(\boldsymbol{x}_s^t) - \nabla f_{i_t}(\boldsymbol{x}_s^{t-1})}_{\text{stochastic gradient}} + \boldsymbol{g}_s^{t-1}$

7: $\quad\quad \boldsymbol{x}_s^{t+1} = \boldsymbol{x}_s^t - \eta \boldsymbol{g}_s^t$

SARAH

Theorem 4 ([Nguyen et al., 2019])

Suppose each f_i is L-smooth. Then SARAH with $\eta \lesssim \frac{1}{L\sqrt{m}}$ obeys

$$\frac{1}{(m+1)S}\sum_{s=1}^{S}\sum_{t=0}^{m}\mathbb{E}\left[\left\|\nabla F(x_s^t)\right\|_2^2\right] \le \frac{2}{\eta(m+1)S}[F(x_0^0) - F(x^*)].$$

SARAH

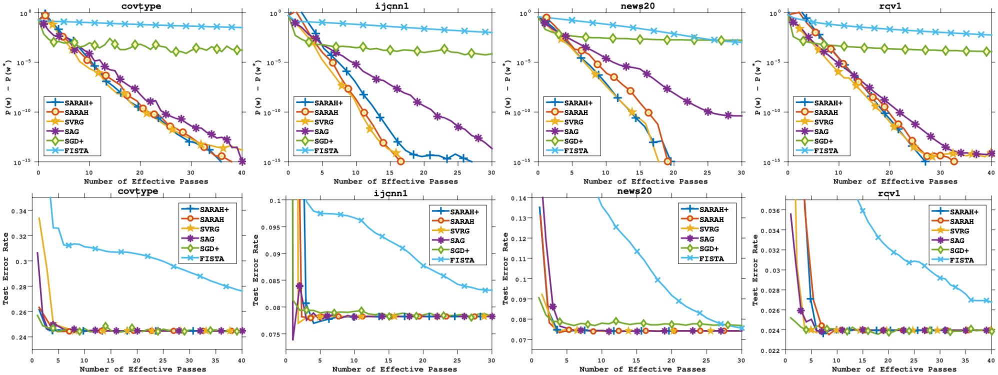


Figure 5: Comparisons of loss residuals $P(w) - P(w^*)$ (top) and test errors (bottom) from different modern stochastic methods on *covtype*, *ijcnn1*, *news20* and *rcv1*.

Brief Survey on Incremental Gradient Algorithms

Incremental gradient descent algorithms were developed to have such characteristics, and hence form an important class of algorithms. A list of few popular incremental algorithms is given below.

Deterministic Incremental Gradient Algorithms

1. Incremental Gradient Descent (IGD) [Bertsekas, 1997]
2. Incremental Aggregated Gradient (IAD) [Blatt et al., 2007]

Stochastic Incremental Gradient Algorithms

1. Stochastic Average Gradient (SAG) [Schmidt et al., 2017]
2. SAGA [Defazio et al., 2014a]

Brief Survey on Incremental Gradient Algorithms

Stochastic Incremental Gradient Algorithms

3. Stochastic Variance Reduced Gradient (SVRG) [Johnson and Zhang, 2013]
4. Semi-Stochastic Gradient Descent (S2GD) [Konečnỳ and Richtárik, 2013]
5. Faster Permutable Incremental Gradient Method (Finito) [Defazio et al., 2014b]
6. Miminization by Incremental Surrogate Optimization (MISO)[Mairal, 2013]
7. Randomized Primal-Dual Gradient (RPDG)[Lan and Zhou, 2018]
8. StochAstic Recursive grAdient algoritHm (SARAH) [Nguyen et al., 2017, 2019]
9. Spider [Fang et al., 2018]
10. Storm [Cutkosky and Orabona, 2019]

References I

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pages 1–9, 2015.

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in neural information processing systems*, pages 1017–1025, 2014.

Peilin Zhao and Tong Zhang. Accelerating minibatch stochastic gradient descent using stratified sampling. *arXiv preprint arXiv:1405.3080*, 2014.

Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

References II

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1953048.2021068`.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8), 2012.

References III

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Timothy Dozat. Incorporating nesterov momentum into adam. 2016.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2613–2621. JMLR. org, 2017.

References IV

Lam M Nguyen, Marten van Dijk, Dzung T Phan, Phuong Ha Nguyen, Tsui-Wei Weng, and Jayant R Kalagnanam. Finite-sum smooth optimization with sarah. *arXiv preprint arXiv:1901.07648*, 2019.

Dimitri P Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM Journal on Optimization*, 7(4):913–926, 1997.

Doron Blatt, Alfred O Hero, and Hillel Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014a.

References V

Jakub Konečnỳ and Peter Richtárik. Semi-stochastic gradient descent methods. *arXiv preprint arXiv:1312.1666*, 2013.

Aaron Defazio, Justin Domke, et al. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pages 1125–1133, 2014b.

Julien Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791, 2013.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, 171(1-2):167–215, 2018.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

References VI

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Alekh Agarwal, Peter L Bartlett, Pradeep Ravikumar, and Martin J Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 5(58):3235–3249, 2012.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley, 1983.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Niao He. Big data optimizaton course, ece236c. URL `http://niaohe.ise.illinois.edu/IE598_2016/index.html`.

References VII

Yuxin Chen. Large-scale optimization for data science, ele522. URL http://www.princeton.edu/~yc5/ele522_optimization/.

Thank You!

Email:qianhui@zju.edu.cn