Stochastic Optimization for Large-scale Machine Learning

Wei-Cheng Chang

Language Technologies Institute
Carnegie Mellon University

September 13, 2016
Outline

1 Introduction

2 Analysis of Stochastic Gradient Methods
   - Preliminaries
   - Two Assumptions and Fundamental Lemmas
   - SG for Strongly Convex Objectives
   - SG for General Objectives

3 Noise Reduction Methods
   - Dynamic Sampling
   - Gradient Aggregation

4 Second-Order Methods
   - Hessian Free methods
   - Stochastic Quasi-Newton Methods
   - Gauss-Newton
   - Diagonal Scaling

5 Conclusions
Expected Risk: Consider prediction function $h$ parameterized by $w \in \mathbb{R}^d$ and some loss function $\ell(\cdot, \cdot)$,

$$R(w) = \int \ell(h(x; w), y) dP(x, y) = \mathbb{E}[\ell(h(x; w), y)].$$

Empirical Risk: In supervised learning, only have access to a set of $n$ I.I.D. samples $\{(x_i, y_i)\} \subseteq \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

$$R_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i; w), y_i).$$

We consider $R_n$ as the objective function from now on. Note that all methods in the subsequent slides can be applied readily when a smooth regularization term.
a sets of samples is presented by a random seed $\xi$. For example, a realization of $\xi$ as a single sample $(x, y)$ or a set of samples $\{(x_i, y_i)\}_{i \in S}$.

refer the loss incurred for a given $(w, \xi)$ as $f(w; \xi)$. That is, $f$ is the composition of loss function $\ell$ and prediction function $h$.

Then we have,

$$R(w) = \mathbb{E}[f(w; \xi)],$$

$$R_n(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w),$$

where $f_i(w) \equiv f(w; \xi_{[i]})$ and $\{\xi_{[i]}\}_{i=1}^{n}$ is a set of realizations of $\xi$ corresponding to a sample set $\{(x_i, y_i)\}$.
Why Stochastic Optimization in Large-scale Machine Learning?

- Batch gradient methods could achieve **linear rate of convergence**

\[ R_n(w_k) - R_n^* \leq O(\rho^k). \]

To obtain \(\epsilon\)-optimality, the total work is proportion to \(n \log(1/\epsilon)\).

- SG could achieve **sublinear rate of convergence in expectation**

\[ \mathbb{E}[R_n(w_k) - R_n^*] = O(1/k). \]

To obtain \(\epsilon\)-optimality, the total work is proportion to \(O(1/\epsilon)\). When \(n\) is large and limited computational time budget, SG is favorable. (More on later slides)

- Due to stochastic nature of SG, it enjoys same rate of convergence for the generalization error \(R - R^*\).
1 Introduction

2 Analysis of Stochastic Gradient Methods
   • Preliminaries
     • Two Assumptions and Fundamental Lemmas
     • SG for Strongly Convex Objectives
     • SG for General Objectives

3 Noise Reduction Methods
   • Dynamic Sampling
   • Gradient Aggregation

4 Second-Order Methods
   • Hesssian Free methods
   • Stochastic Quasi-Newton Methods
   • Gauss-Newton
   • Diagonal Scaling

5 Conclusions
Consider optimization problem

\[ f^* = \min_x f(x), \]

linear rate of convergence means, for some \( c \in (0, 1) \),

\[ f(x_k) - f^* \leq c(f(x_{k-1}) - f^*) \leq \ldots \leq c^k(f(x_0) - f^*) \leq \epsilon \]

to achieve \( \epsilon \)-optimality, the number of iterations need at most

\[ k \leq \frac{\log((f(x_0) - f^*)/\epsilon)}{\log(1/c)} = O\left( \frac{\log(1/\epsilon)}{\log(1/c)} \right) \]
Objective function $F : \mathbb{R}^d \rightarrow \mathbb{R}$ represents

$$R(w) = \mathbb{E}[f(w; \xi)] \text{ or } R_n(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

SG updates:

$$w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k).$$

Our analyses cover the choices

$$g(w_k, \xi_k) = \begin{cases} \\
\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n}\n
$$
$$

where $H_k$ is PSD, and $n_k$ is the mini-batch size.
Some Intuitions of SG

- $\{\xi_k\}$ is a sequence of independent random variables.
- Sample uniformly from training set, SG optimizes the empirical risk $F = R_n$.
- Picking samples according to $P$, SG optimizes the expected risk $F = R$.
- Our goal is to show the expected angle between $g(w_k, \xi_k)$ and $\nabla F(w_k)$ is sufficiently close.
1. Introduction

2. Analysis of Stochastic Gradient Methods
   - Preliminaries
   - Two Assumptions and Fundamental Lemmas
     - SG for Strongly Convex Objectives
     - SG for General Objectives

3. Noise Reduction Methods
   - Dynamic Sampling
   - Gradient Aggregation

4. Second-Order Methods
   - Hessian Free methods
   - Stochastic Quasi-Newton Methods
   - Gauss-Newton
   - Diagonal Scaling

5. Conclusions
Assumption 4.1 (Lipschitz-continuous objective gradients)

$F$ is continuously differentiable and the gradient function of $F$, namely $\nabla F$, is Lipschitz-continuous with Lipschitz constant $L > 0$, i.e.,

$$\|\nabla F(w) - \nabla F(\bar{w})\|_2 \leq L \|w - \bar{w}\|_2 \quad \forall \{w, \bar{w}\} \subset \mathbb{R}^d$$

This assumption ensures gradient of $F$ does not change arbitrary quickly w.r.t the parameter $w$. From this assumption, we could then derive

$$F(w) \leq F(\bar{w}) + \nabla F(\bar{w})^T(w - \bar{w}) + \frac{1}{2}L\|w - \bar{w}\|_2^2$$  \hspace{1cm} (1)
Proof of equation (1)

Define $g(t) = F(\bar{w} + t(w - \bar{w}))$ and notice

$$g'(t) = \nabla F(\bar{w} + t(w - \bar{w}))^T (w - \bar{w}).$$

$$g(1) = g(0) + \int_0^1 g'(t) \, dt$$

$$= F(\bar{w}) + \int_0^1 \nabla F(\bar{w} + t(w - \bar{w}))^T (w - \bar{w}) \, dt$$

$$= F(\bar{w}) + \nabla F(\bar{w})^T (w - \bar{w}) + \int_0^1 \left[ \nabla F(\bar{w} + t(w - \bar{w})) - \nabla F(\bar{w}) \right]^T (w - \bar{w}) \, dt$$

$$\leq F(\bar{w}) + \nabla F(\bar{w})^T (w - \bar{w}) + \int_0^1 L \|t(w - \bar{w})\|_2 \|w - \bar{w}\|_2 \, dt$$

$$= F(\bar{w}) + \nabla F(\bar{w})^T (w - \bar{w}) + \frac{1}{2} L \|w - \bar{w}\|_2^2.$$
Bound expected function decrease by gradient information

Lemma 4.2

Under Assumption 1, the iterates of SG satisfy the following inequality

\[
\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k} [g(w_k, \xi_k)] \\
+ \frac{1}{2} \alpha_k^2 L \mathbb{E}_{\xi_k} [\|g(w_k, \xi_k)\|^2].
\] (2)

Proof: By Assumption 4.1, the iterates by SG satisfy

\[
F(w_{k+1}) - F(w_k) \leq \nabla F(w_k)^T (w_{k+1} - w_k) + \frac{1}{2} L \|w_{k+1} - w_k\|^2 \\
\leq -\alpha_k \nabla F(w_k)^T g(w_k, \xi_k) + \frac{1}{2} \alpha_k^2 L \|g(w_k, \xi_k)\|^2.
\]

Taking expectations w.r.t. the distribution of \( \xi_k \), we obtain the desired results.
Intuitions of Lemma 4.2

- RHS of (2) depends on the first and second moment of $g(w_k, \xi_k)$.
- If $g(w_k, \xi_k)$ is an unbiased estimate of $\nabla F(w_k)$, then

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L\mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2].$$

- Convergence is guarantee if RHS of (2) is be bounded.
- We need to restricted the variance of $g(w_k, \xi_k)$, i.e.,

$$\nabla_{\xi_k}[g(w_k, \xi_k)] := \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] - \|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2^2. \quad (3)$$
Assumption 4.3 (First and Second Moment limits)

The objective function and SG satisfy the following

1. The sequence of iterates \( \{w_k\} \) is contained in an open set over which \( F \) is bounded below by \( F_{\inf} \).

2. There exist scalars \( \mu_G \geq \mu > 0 \) such that

\[
\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \geq \mu \| \nabla F(w_k) \|^2_2, \\
\| \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \| \leq \mu G \| \nabla F(w_k) \|_2.
\]  

3. There exist scalars \( M > 0 \) and \( M_V > 0 \) such that

\[
\nabla_{\xi_k}[g(w_k, \xi_k)] \leq M + M_v \| \nabla F(w_k) \|^2_2.
\]
The properties in Assumption 4.3-(2) immediately holds if \( g(w_k, \xi_k) \) is an unbiased estimate of \( F(w_k) \), and are maintained if such unbiased estimate is multiplied by a PSD matrix \( H_k \).

Combing the definition of variance (3) and Assumption 4.3,

\[
\mathbb{E}_{\xi_k} [\| g(w_k, \xi_k) \|^2_2] \leq M + M_G \| \nabla F(w_k) \|, \tag{7}
\]

where \( M_G := M_V + \mu_G^2 \geq \mu > 0 \).
Lemma 4.4

Under Assumption 4.1 and 4.3, the iterate of SG satisfy the following inequality

$$
\mathbb{E}[F(w_{k+1})] - F(w_k) \leq -\left(\mu - \frac{1}{2} \alpha_k LM_G \right) \alpha_k \|\nabla F(w_k)\|^2 + \frac{1}{2} \alpha^2_k LM.
$$

Proof. By Lemma 1 and (4), it follows that

$$
\mathbb{E}[F(w_{k+1})] - F(w_k) \leq -\mu \alpha_k \|\nabla F(w_k)\|^2 + \frac{1}{2} \alpha^2_k L \mathbb{E}[\|g(w_k, \xi_k)\|^2].
$$

Together with (7), it yields the desired results.
Outline

1 Introduction

2 Analysis of Stochastic Gradient Methods
   - Preliminaries
   - Two Assumptions and Fundamental Lemmas
   - SG for Strongly Convex Objectives
   - SG for General Objectives

3 Noise Reduction Methods
   - Dynamic Sampling
   - Gradient Aggregation

4 Second-Order Methods
   - Hessian Free methods
   - Stochastic Quasi-Newton Methods
   - Gauss-Newton
   - Diagonal Scaling

5 Conclusions
Assumption 4.5 (Strong Convexity)

The objective function $F$ is strongly convex if there exist a constant $c > 0$ such that

$$
\nabla^2 F(w) \geq cI \quad \forall w \in \mathbb{R}^d.
$$

Strong convexity has several interesting consequences.

- Better lower bound on $F(w)$

  $$
  F(w) \geq F(\tilde{w}) + \nabla F(\tilde{w})^T (w - \tilde{w}) + \frac{c}{2} \|w - \tilde{w}\|^2.
  $$

- Bound the optimality gap $F(w) - F(w^*)$

  $$
  2c(F(w) - F(w^*)) \leq \|\nabla F(w)\|^2 \quad \forall w \in \mathbb{R}^d
  $$

(8)
Strong Convexity (Cont’d)

\[
F(\bar{w}) \geq F(w) + \nabla F(w)^T (\bar{w} - w) + \frac{c}{2} \| \bar{w} - w \|^2 \\
\geq F(w) + \nabla F(w)^T (\tilde{w} - w) + \frac{c}{2} \| \tilde{w} - w \|^2 \\
= F(w) - \frac{1}{2c} \| \nabla F(w) \|^2,
\]

where \( \tilde{w} = w - (1/c) \| \) is the minimizer of the quadratic model. This holds for any \( \bar{w} \in \mathbb{R}^d \). Therefore,

\[
F(w^*) \geq F(w) - \frac{1}{2c} \| \nabla F(w) \|^2.
\]
Theorem 4.6 (Strongly Convex Objective, Fixed Stepsizes)

Under assumption 4.1, 4.3, 4.5, suppose that the SG method is run with a fixed stepsize $\alpha_k = \bar{\alpha}$, satisfying

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}. \quad (9)$$

Then, the expected optimality gap satisfies

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\bar{\alpha}LM}{2c\mu} + (1 - \bar{\alpha}c\mu)^{k-1}\left(F(w_1) - F_* - \frac{\bar{\alpha}LM}{2c\mu}\right) \quad (10)$$

$$\xrightarrow{k \to \infty} \frac{\bar{\alpha}LM}{2c\mu}, \quad (11)$$

where $\mathbb{E}[F(w_k)] = \mathbb{E}_{\xi_1}\mathbb{E}_{\xi_2} \ldots \mathbb{E}_{\xi_{k-1}}[F(w_k)]$. 
Intuition of Theorem 4.6

- If there’s no noise ($M = 0$), Theorem 4.6 becomes linear rate of convergence. Reduce to standard result of full gradient.
- Need $\bar{\alpha}$ to be small.
**Theorem 4.7 (Strongly Convex Objective, Diminishing Stepsizes)**

Under assumption 4.1, 4.3, 4.5, suppose that the SG method is run with a stepsize sequence, such that

\[ \alpha_k = \frac{\beta}{\gamma + k} \]

for some \( \beta > \frac{1}{c\mu} \) and \( \gamma > 0 \) such that

\[ \alpha_1 \leq \frac{\mu}{LM_G}. \quad (12) \]

Then, the expected optimality gap satisfies

\[ \mathbb{E}[F(w_k) - F^*] \leq \frac{\nu}{\gamma + k}, \quad (13) \]

where

\[ \nu := \max \left\{ \frac{\beta^2 LM}{2(\beta c\mu - 1)}, (\gamma + 1)(F(w_1) - F^*) \right\}. \quad (14) \]
Inuition of Theorem 4.7

- For constant stepsize, $\bar{\alpha}$ does not depend on $c$. For diminishing stepsizes, initial $\alpha_1$ depends on $c$.
- For constant stepsize, initial point eventually does not matter. For diminishing stepsizes, initial point matters.
- The effect of mini-batch is "subtle"
1 Introduction

2 Analysis of Stochastic Gradient Methods
   - Preliminaries
   - Two Assumptions and Fundamental Lemmas
   - SG for Strongly Convex Objectives
   - SG for General Objectives

3 Noise Reduction Methods
   - Dynamic Sampling
   - Gradient Aggregation

4 Second-Order Methods
   - Hessian Free methods
   - Stochastic Quasi-Newton Methods
   - Gauss-Newton
   - Diagonal Scaling

5 Conclusions
Theorem 4.8 (Nonconvex Objective, Fixed Stepsize)

Under assumption 4.1, 4.3, suppose that the SG method is run with a fixed stepsize $\alpha_k = \bar{\alpha}$, satisfying

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}. \quad (15)$$

Then, the expected sum-of-squares and average-squared gradients of $F$ satisfies

$$\mathbb{E} \left[ \sum_{k=1}^{K} \|\nabla F(w_k)\|^2 \right] \leq \frac{K \bar{\alpha} LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{\mu \bar{\alpha}} \quad (16)$$

$$\mathbb{E} \left[ \frac{1}{K} \sum_{k=1}^{K} \|\nabla F(w_k)\|^2 \right] \leq \frac{\bar{\alpha} LM}{\mu} + \frac{2(F(w_1) - F_{inf})}{K \mu \bar{\alpha}} \quad (17)$$

$$K \to \infty \quad \frac{\bar{\alpha} LM}{\mu}. \quad (18)$$
Theorem 4.9 (Nonconvex Objective, Diminishing Stepsizes)

Under Assumption 4.1 and 4.3, suppose SG is run with a stepsize sequence satisfying $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, with $A_K = \sum_{k=1}^{K} \alpha_k$,

$$\mathbb{E}\left[\sum_{k=1}^{K} \alpha_k \| \nabla F(w_k) \|^2 \right] < \infty$$  \hspace{1cm} (19)

$$\mathbb{E}\left[\frac{1}{A_K} \sum_{k=1}^{K} \alpha_k \| \nabla F(w_k) \|^2 \right] \xrightarrow{K \to \infty} 0.$$  \hspace{1cm} (20)
Outline

1 Introduction

2 Analysis of Stochastic Gradient Methods
   - Preliminaries
   - Two Assumptions and Fundamental Lemmas
   - SG for Strongly Convex Objectives
   - SG for General Objectives

3 Noise Reduction Methods
   - Dynamic Sampling
   - Gradient Aggregation

4 Second-Order Methods
   - Hessian Free methods
   - Stochastic Quasi-Newton Methods
   - Gauss-Newton
   - Diagonal Scaling

5 Conclusions
Recall that Lemma 4.2, we have

$$
\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k} [g(w_k, \xi_k)] \\
+ \frac{L}{2} \alpha_k^2 \mathbb{E}_{\xi_k} \|g(w_k, \xi_k)\|_2^2.
$$

If $g()$ is a descent direction in expectation and we could decrease $\mathbb{E}_{\xi_k} \|g(w_k, \xi_k)\|_2^2$ fast enough, then good convergence rate is still possible.
Theorem 5.1 (Strongly Convex Objective, Noise Reduction)

Suppose that Assumption 4.1, 4.3 and 4.5 hold, and that

\[ \nabla_{\xi_k} [g(w_k, \xi_k)] \leq M \zeta^{k-1}. \]  \hspace{1cm} (21)

In addition, suppose SG is run with a fixed stepsize, \( \alpha_k = \bar{\alpha} \) such that

\[ 0 < \bar{\alpha} \leq \min\left\{ \frac{\mu}{L\mu_G^2}, \frac{1}{c\mu} \right\}. \]

Then the expected optimality gap satisfies

\[ \mathbb{E}\left[F(w_k) - F_*\right] \leq \omega \rho^{k-1} \]

where

\[ \omega := \max\left\{ \frac{\bar{\alpha}LM}{c\mu}, F(w_1) - F_* \right\} \text{ and } \rho := \max\left\{ 1 - \bar{\alpha}c\mu/2, \zeta \right\}. \]
Consider the iteration

\[ w_{k+1} \leftarrow w_k - \frac{\bar{\alpha}}{n_k} \sum_{i \in S_k} \nabla f(w_k, \xi_k, i) \]

with \( n_k := |S_k| = \lceil \tau^{k-1} \rceil \) for some \( \tau > 1 \). Then this update fall under the class of methods in Theorem 5.1, because

\[
\nabla_{\xi_k} g(w_k, \xi_k) \leq \frac{\nabla_{\xi_k} f(w_k; \xi_k, i)}{n_k} \leq \frac{M}{n_k},
\]
Outline

1 Introduction

2 Analysis of Stochastic Gradient Methods
   • Preliminaries
   • Two Assumptions and Fundamental Lemmas
   • SG for Strongly Convex Objectives
   • SG for General Objectives

3 Noise Reduction Methods
   • Dynamic Sampling
   • Gradient Aggregation

4 Second-Order Methods
   • Hessean Free methods
   • Stochastic Quasi-Newton Methods
   • Gauss-Newton
   • Diagonal Scaling

5 Conclusions
Rather than increasing more new stochastic gradient, is it possible to achieve a lower variance by reusing and/or revising previous computed stochastic gradient information?

- Stochastic variance reduction gradient (SVRG)
- Stochastic average gradient (SAG)
- Iterative averaging

\[
\begin{align*}
    w_{k+1} &\leftarrow w_k - \alpha_k g(w_k, \xi_k) \\
    \tilde{w}_{k+1} &\leftarrow \frac{1}{k+1} \sum_{j=1}^{k+1} w_j.
\end{align*}
\]
Outline

1. Introduction
2. Analysis of Stochastic Gradient Methods
   - Preliminaries
   - Two Assumptions and Fundamental Lemmas
   - SG for Strongly Convex Objectives
   - SG for General Objectives
3. Noise Reduction Methods
   - Dynamic Sampling
   - Gradient Aggregation
4. Second-Order Methods
   - Hessian Free methods
   - Stochastic Quasi-Newton Methods
   - Gauss-Newton
   - Diagonal Scaling
5. Conclusions
Preliminaries

- To address adverse effects of high nonlinearity and ill-condition of objective function
- At each iteration, solve a quadratic model

\[ q_k(s) = F(w_k) + \nabla F(w_k)^T s + \frac{1}{2} s^T H_k s \]

- The Hessian matrix \( H_k \in \mathbb{R}^{d \times d} \) (i.e., \( \nabla^2 F(w_k) \)) could be very large, need to solve the following linear system

\[ H_k s = -\nabla F(w_k) \]  \( (22) \)

- The update

\[ w_{k+1} \leftarrow w_k + \alpha_k s_k \]
When $d$ is large (especially in deep learning), impossible to explicitly compute $H$.

Apply Conjugate Gradient (CG) method to iteratively solve the linear system (22).

Only need to compute matrix-vector product $Hs$, as expensive as computing gradient.

CG need at most $d$ iterations to solve the linear system.

This approach is still not possible for deep neural nets...
Inexact Newton suggests it is tolerant to certain degree of noise in the Hessian estimate than it is to the noise in the gradient estimate.

One could employ subsampled Hessian-free inexact Newton

$$\nabla^2 f_{S_k^H}(w_k; \xi_k^H)s = -\nabla f_{S_k}(w_k; \xi_k)$$

where $S_k^H \subset S_k$.

gradient sample size $|S_k|$ need to be large enough so that taking subsamples for the Hessian estimate is sensible.
Approximate Hessian using only gradient information

BFGS is one of the most success of Quasi-Newton, by defining

\[ s_k := w_{k+1} - w_k \] and \[ v_k := \nabla F(w_{k+1}) - \nabla F(w_k), \]

the update rule of BFGS is

\[ w_{k+1} \leftarrow w_k - H_k \nabla F(w_k) \]

\[ H_{k+1} \leftarrow \left(1 - \frac{v_k s_k^T}{s_k^T v_k}\right)^T H_k \left(1 - \frac{v_k s_k^T}{s_k^T v_k}\right) + \frac{s_k s_K^T}{s_k^T v_k}, \]

where \( H_k \) is a PSD approximation of \( (\nabla^2 F(w_k))^{-1} \).
This update ensures $H_{k+1}^{-1} s_k = v_k$ holds, meaning that a second-order Taylor expansion is satisfied along the most recent displacement.

- Enjoy a local superlinear rate of convergence.

- $H_{k+1}$ are dense, even if the exact Hessians are sparse.

- L-BFGS to rescue: don’t explicitly form $H_k$, but only compute $H_k \nabla F(w_k)$ by a sequence of displacement pairs $\{(s_k, v_k)\}$.

- L-BFGS only achieves linear rate of convergence though...
How about Stochastic L-BFGS?

- Iteration update: \( w_{k+1} \leftarrow w_k - \alpha_k H_k g(w_k, \xi_k) \)
- Stochastic L-BFGS could only have sublinear rate as SG. What’s the benefit? improving the constant, from dependent to Hessian to independent of Hessian!
- each iteration costs \( O(4md) \), if \( m = 5 \), the 20 times greater than stochastic gradient estimation. This is not big problem in mini-batch SG though.
- the effects of even a single bad update may linger for numerous iterations.
Replace gradient with stochastic gradient

\[ s_k := w_{k+1} - w_k \quad \text{and} \quad v_k := \nabla f_{S_k}(w_{k+1}, \xi_k) - \nabla f_{S_k}(w_k, \xi_k) \]

- Note that the use of same seed in two gradient estimate.
- The inverse Hessian approximation with ever step may not be warranted when using small sample size \( S_k \).
- Alternatively, since \( \nabla F(w_{k+1}) - \nabla F(w_k) \approx \nabla F(w_k)^2(w_{k+1} - w_k) \), define

\[ v_k := \nabla^2 f_{S_k}(w_k; \xi^H_S) s_k, \]

where \( \nabla^2 f_{S_k}(w_k; \xi^H_S) \) is a subsampled hessian.
Outline

1 Introduction

2 Analysis of Stochastic Gradient Methods
   - Preliminaries
   - Two Assumptions and Fundamental Lemmas
   - SG for Strongly Convex Objectives
   - SG for General Objectives

3 Noise Reduction Methods
   - Dynamic Sampling
   - Gradient Aggregation

4 Second-Order Methods
   - Hessian Free methods
   - Stochastic Quasi-Newton Methods
     - Gauss-Newton
   - Diagonal Scaling

5 Conclusions
**Intuition**

- Construct an approximation to the Hessian using only first-order information.
- Always PSD, even when Hessian is not (useful for deep learning).
- Ignores second-order interaction between elements of the parameter vector $w$. 
Consider the following problem

\[ f(w; \xi) = \ell(h(x_\xi; w), y_\xi) = \frac{1}{2} \| h(x_\xi; w) - y_\xi \|^2_2. \]

Gauss-Newton approximation is an affine approximation of the prediction function inside the quadratic loss function,

\[ h(\xi_i; w) \approx h(x_\xi; w_k) + J_h(w_k; \xi)(w - w_k) \]

where \( J(\cdot; \xi) \) represent the Jacobian of \( h(\xi; \cdot) \) w.r.t. \( w \). This leads to

\[
\begin{align*}
    f(w; \xi) &\approx \frac{1}{2} \| h(x_\xi; w_k) + J_h(w_k; \xi)(w - w_k) - y_\xi \|^2_2 \\
    &= \frac{1}{2} \| h(x_\xi; w) - y_\xi \|^2_2 + (h(x_\xi; w_k) - y_\xi)^T J_h(w_k; \xi)(w - w_k) \\
    &\quad + \frac{1}{2} (w - w_k)^T J_h(w_k; \xi)^T J_h(w_k; \xi)(w - w_k)
\end{align*}
\]
We could replace subsampled Hessian with Gauss-Newton matrix

\[
G_{SH_k}(w_k; \xi_k^H) = \frac{1}{|S_k^H|} \sum_{i \in S_k^H} J_h(w_k; \xi_{k,i})^T J_h(w_i; \xi_{k,i})
\]  \hspace{1cm} (23)

For least square loss function, Gauss-Newton is always PSD, so it could serve as another approximation of hessian in stochastic L-BFGS and Inexact Newton method.

It could also be generalized to other loss functions.

\[
G_{SH_k}(w_k; \xi_k^H) = \frac{1}{|S_k^H|} \sum_{i \in S_k^H} J_h(w_k; \xi_{k,i})^T H(\ell)(w_k; \xi_{k,i}) J_h(w_i; \xi_{k,i}),
\]

where \( H(\ell)(w_k; \xi) = \frac{\partial^2}{\partial h^2} \ell(h(x_\xi; w), y_\xi). \)
Outline

1. Introduction
2. Analysis of Stochastic Gradient Methods
   - Preliminaries
   - Two Assumptions and Fundamental Lemmas
   - SG for Strongly Convex Objectives
   - SG for General Objectives
3. Noise Reduction Methods
   - Dynamic Sampling
   - Gradient Aggregation
4. Second-Order Methods
   - Hessian Free methods
   - Stochastic Quasi-Newton Methods
   - Gauss-Newton
   - Diagonal Scaling
5. Conclusions
Diagonal Scaling Intuition

- Iteration update:
  \[ w_{k+1} \leftarrow w_k - \alpha_k D_k g(w_k, \xi_k) \]

- \( D_k^{-1} \approx \text{diag(Hessian)} \) or \( \text{diag(} \text{Guass-Newton)} \)

- \( D_k^{-1} \approx \text{running average/sum of gradient component.} \)
RMSPROP estimates the average magnitude of each element of the stochastic gradient vector \(g(w_k, \xi_k)\) by maintaining the running averages

\[
R_k = (1 - \lambda)R_{k-1} + \lambda (g(w_k, \xi_k) \circ g(w_k, \xi_k))
\]

\[
w_{k+1} = w_k - \frac{\alpha}{\sqrt{R_k + \mu}} \circ g(w_k, \xi_k)
\]

where \(\circ\) represents element-wise product.

ADAGRAD just replaced the running average by a sum

\[
R_k = R_{k-1} + [g(w_k, \xi_k) \circ g(w_k, \xi_k)]
\]
RMSPROP and ADAGRAD are considered the state-of-the-art stochastic optimization method in practical use of deep learning model.

Note that they are just approximating the diagonal value of Hessian.

There’s still big room for stochastic second order method to be played around with the deep learning model!