Summary of Review Paper: Optimization Methods for Large-Scale Machine Learning

Muthu Chidambaram

Department of Computer Science, University of Virginia

https://qdata.github.io/deep2Read/

Introduction

- **Authors:** Leon Bouttou, Frank E. Curtis, Jorge Nocedal
- Overview of optimization methods
- Characterization of large-scale machine learning as a distinctive setting
- Research directions for next generation of optimization methods

Stochastic vs Batch Gradient Methods

- Stochastic Gradient Descent
  - Formulated as: $w_{k+1} \leftarrow w_k - \alpha_k \nabla f_{i_k}(w_k)$
  - Uses information more efficiently
  - Computationally less expensive
- Batch Gradient Descent
  - Formulated as: $w_{k+1} \leftarrow w_k - \alpha_k \nabla R_n(w_k) = w_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(w_k)$
  - Better performance over large number of epochs
  - Less noisy

Notation

- f: composition of loss and prediction functions
- ξ: random sample or set of samples from data
- w: parameters of prediction function
- f_i: loss with respect to a single sample

SGD Analysis: Lipschitz Continuous

Assumption 4.1 (Lipschitz-continuous objective gradients). *The objective function* $F : \mathbb{R}^d \to \mathbb{R}$ *is continuously differentiable and the gradient function of* F, *namely,* $\nabla F : \mathbb{R}^d \to \mathbb{R}^d$, *is Lipschitz continuous with Lipschitz constant* $L > 0$, *i.e.,*

$$\|\nabla F(w) - \nabla F(\overline{w})\|_2 \leq L\|w - \overline{w}\|_2 \ \ \text{for all} \ \ \{w, \overline{w}\} \subset \mathbb{R}^d.$$

Lemma 4.2. *Under Assumption 4.1, the iterates of SG (Algorithm 4.1) satisfy the following inequality for all* $k \in \mathbb{N}$:

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \tfrac{1}{2}\alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2]. \tag{4.4}$$

SGD Analysis: Restrictions on Moments

Assumption 4.3 (First and second moment limits). *The objective function and SG (Algorithm 4.1) satisfy the following:*

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

(b) There exist scalars $\mu_G \geq \mu > 0$ *such that, for all* $k \in \mathbb{N}$,

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \geq \mu \|\nabla F(w_k)\|_2^2 \quad \text{and} \tag{4.7a}$$

$$\|\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2. \tag{4.7b}$$

(c) There exist scalars $M \geq 0$ *and* $M_V \geq 0$ *such that, for all* $k \in \mathbb{N}$,

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] \leq M + M_V \|\nabla F(w_k)\|_2^2. \tag{4.8}$$

SGD Analysis: Strongly Convex (Fixed Stepsize)

Assumption 4.5 (Strong convexity). *The objective function $F : \mathbb{R}^d \to \mathbb{R}$ is strongly convex in that there exists a constant $c > 0$ such that*

$$F(\overline{w}) \geq F(w) + \nabla F(w)^T(\overline{w} - w) + \tfrac{1}{2}c\|\overline{w} - w\|_2^2 \quad \text{for all} \quad (\overline{w}, w) \in \mathbb{R}^d \times \mathbb{R}^d. \tag{4.11}$$

Hence, F has a unique minimizer, denoted as $w_* \in \mathbb{R}^d$ with $F_* := F(w_*)$.

Theorem 4.6 (Strongly Convex Objective, Fixed Stepsize). *Under Assumptions 4.1, 4.3, and 4.5 (with $F_{\text{inf}} = F_*$), suppose that the SG method (Algorithm 4.1) is run with a fixed stepsize, $\alpha_k = \bar{\alpha}$ for all $k \in \mathbb{N}$, satisfying*

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}. \tag{4.13}$$

Then, the expected optimality gap satisfies the following inequality for all $k \in \mathbb{N}$:

$$\begin{aligned}\mathbb{E}[F(w_k) - F_*] &\leq \frac{\bar{\alpha}LM}{2c\mu} + (1 - \bar{\alpha}c\mu)^{k-1}\left(F(w_1) - F_* - \frac{\bar{\alpha}LM}{2c\mu}\right) \\ &\xrightarrow{k\to\infty} \frac{\bar{\alpha}LM}{2c\mu}.\end{aligned} \tag{4.14}$$

SGD Analysis: Strongly Convex (Diminishing Stepsize)

Theorem 4.7 (Strongly Convex Objective, Diminishing Stepsizes). *Under Assumptions 4.1, 4.3, and 4.5 (with $F_{\text{inf}} = F_*$), suppose that the SG method (Algorithm 4.1) is run with a stepsize sequence such that, for all $k \in \mathbb{N}$,*

$$\alpha_k = \frac{\beta}{\gamma + k} \quad \text{for some} \quad \beta > \frac{1}{c\mu} \quad \text{and} \quad \gamma > 0 \quad \text{such that} \quad \alpha_1 \leq \frac{\mu}{LM_G}. \tag{4.18}$$

Then, for all $k \in \mathbb{N}$, the expected optimality gap satisfies

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\nu}{\gamma + k}, \tag{4.19}$$

where

$$\nu := \max\left\{\frac{\beta^2 LM}{2(\beta c\mu - 1)}, (\gamma + 1)(F(w_1) - F_*)\right\}. \tag{4.20}$$

Roles of Assumptions

- Strong Convexity
  - Key for ensuring O(1/k) convergence
- Initialization
  - Can be used to decrease the prominence of initial gap in decreasing stepsize optimization

SGD Analysis: General Objectives

Theorem 4.8 (Nonconvex Objective, Fixed Stepsize). *Under Assumptions 4.1 and 4.3, suppose that the SG method (Algorithm 4.1) is run with a fixed stepsize, $\alpha_k = \bar{\alpha}$ for all $k \in \mathbb{N}$, satisfying*

$$0 < \bar{\alpha} \leq \frac{\mu}{LM_G}. \tag{4.25}$$

Then, the expected sum-of-squares and average-squared gradients of F corresponding to the SG iterates satisfy the following inequalities for all $K \in \mathbb{N}$:

$$\mathbb{E}\left[\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \leq \frac{K\bar{\alpha}LM}{\mu} + \frac{2(F(w_1) - F_{\text{inf}})}{\mu\bar{\alpha}} \tag{4.26a}$$

$$\text{and therefore } \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(w_k)\|_2^2\right] \leq \frac{\bar{\alpha}LM}{\mu} + \frac{2(F(w_1) - F_{\text{inf}})}{K\mu\bar{\alpha}} \tag{4.26b}$$

$$\xrightarrow{K\to\infty} \frac{\bar{\alpha}LM}{\mu}.$$

SGD Analysis: General Objectives

Theorem 4.10 (Nonconvex Objective, Diminishing Stepsizes). *Under Assumptions 4.1 and 4.3, suppose that the SG method (Algorithm 4.1) is run with a stepsize sequence satisfying* (4.17). *Then, with* $A_K := \sum_{k=1}^{K} \alpha_k$,

$$\mathbb{E}\left[\sum_{k=1}^{K} \alpha_k \|\nabla F(w_k)\|_2^2\right] < \infty \tag{4.28a}$$

$$\text{and therefore } \mathbb{E}\left[\frac{1}{A_K}\sum_{k=1}^{K} \alpha_k \|\nabla F(w_k)\|_2^2\right] \xrightarrow{K\to\infty} 0. \tag{4.28b}$$

Complexity for Large-Scale Learning

- Consider infinite supply of training examples
- Batch gradient descent increases linearly
- SGD is independent of training examples

		Batch	Stochastic
$\mathcal{T}(n,\epsilon)$	$\sim$	$n \log\left(\frac{1}{\epsilon}\right)$	$\frac{1}{\epsilon}$
$\mathcal{E}^*$	$\sim$	$\frac{\log(\mathcal{T}_{\max})}{\mathcal{T}_{\max}} + \frac{1}{\mathcal{T}_{\max}}$	$\frac{1}{\mathcal{T}_{\max}}$

SGD Noise Reduction Methods

- Dynamic sampling
  - Minibatches
- Gradient aggregation
  - Store previous gradients
- Iterate averaging
  - Average of iterated values

SGD Noise Reduction Behavior

Theorem 5.1 (Strongly Convex Objective, Noise Reduction). *Suppose that Assumptions 4.1, 4.3, and 4.5 (with $F_{\inf} = F_*$) hold, but with (4.8) refined to the existence of constants $M \geq 0$ and $\zeta \in (0,1)$ such that, for all $k \in \mathbb{N}$,*

$$\mathbb{V}_{\xi_k}[g(w_k, \xi_k)] \leq M\zeta^{k-1}. \tag{5.1}$$

In addition, suppose that the SG method (Algorithm 4.1) is run with a fixed stepsize, $\alpha_k = \bar{\alpha}$ for all $k \in \mathbb{N}$, satisfying

$$0 < \bar{\alpha} \leq \min\left\{\frac{\mu}{L\mu_G^2}, \frac{1}{c\mu}\right\}. \tag{5.2}$$

Then, for all $k \in \mathbb{N}$, the expected optimality gap satisfies

$$\mathbb{E}[F(w_k) - F_*] \leq \omega\rho^{k-1}, \tag{5.3}$$

where

$$\omega := \max\{\tfrac{\bar{\alpha}LM}{c\mu}, F(w_1) - F_*\} \tag{5.4a}$$

$$\textit{and} \;\; \rho := \max\{1 - \tfrac{\bar{\alpha}c\mu}{2}, \zeta\} < 1. \tag{5.4b}$$

SGD Dynamic Sampling

- Increasing minibatch size geometrically guarantees linear convergence
- Practical implementations: adaptive sampling
  - Not tried extensively in ML

SGD Gradient Aggregation

- Stochastic Variance Reduced Gradient (SVRG)
  - Start with batch update and use to correct bias in SGD
- SAGA
  - Uses average of previous gradients to unbias SGD

$$\tilde{g}_j \leftarrow \nabla f_{i_j}(\tilde{w}_j) - (\nabla f_{i_j}(w_k) - \nabla R_n(w_k))$$

$$g_k \leftarrow \nabla f_j(w_k) - \nabla f_j(w_{[j]}) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_{[i]})$$

SGD Iterate Averaging

- Take average of computed parameters to reduce noise

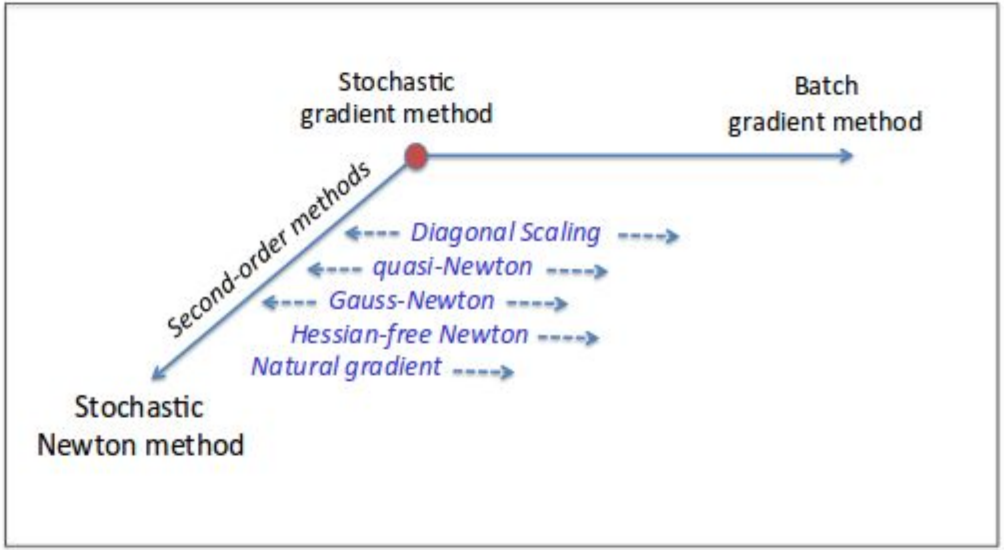
$$w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k)$$

$$\text{and} \quad \tilde{w}_{k+1} \leftarrow \frac{1}{k+1} \sum_{j=1}^{k+1} w_j,$$

Second-Order Methods

- Motivation: SGD not scale invariant
- Hessian-free Newton Method
  - Uses second-order information
- Quasi-Newton and Gauss-Newton Methods
  - Mimic Newton method using sequence of first order information
- Natural Gradient
  - Defines search direction in the space of realizable distributions

Second-Order Method Overview



Hessian-Free Inexact Newton Methods

- Solve Newton system with CG instead of matrix factorization
  - Only requires Hessian vector products
    - Similar to kernel trick

Example 6.1. *Consider the function of the parameter vector* $w = (w_1, w_2)$ *given by* $F(w) = \exp(w_1 w_2)$. *Let us define, for any* $d \in \mathbb{R}^2$, *the function*

$$\phi(w; d) = \nabla F(w)^T d = w_2 \exp(w_1 w_2) d_1 + w_1 \exp(w_1 w_2) d_2.$$

Computing the gradient of ϕ *with respect to* w, *we have*

$$\nabla_w \phi(w; d) = \nabla^2 F(w) d = \begin{bmatrix} w_2^2 \exp(w_1 w_2) d_1 + (\exp(w_1 w_2) + w_1 w_2 \exp(w_1 w_2)) d_2 \\ (\exp(w_1 w_2) + w_1 w_2 \exp(w_1 w_2)) d_1 + w_1^2 \exp(w_1 w_2) d_2 \end{bmatrix}.$$

Subsampled Hessian-Free Newton Methods

Algorithm 6.1 Subsampled Hessian-Free Inexact Newton Method

Choose an initial iterate w_1.

Choose constants $\rho \in (0,1)$, $\gamma \in (0,1)$, $\eta \in (0,1)$, and $\max_{cg} \in \mathbb{N}$.

for $k = 1, 2, \dots$ **do**

Generate realizations of ξ_k and ξ_k^H corresponding to $\mathcal{S}_k^H \subseteq \mathcal{S}_k$.

Compute s_k by applying Hessian-free CG to solve

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

until $\max_{cg}$ iterations have been performed or a trial solution yields

$$\|r_k\|_2 := \|\nabla^2 f_{\mathcal{S}_k^H}(w_k; \xi_k^H) s + \nabla f_{\mathcal{S}_k}(w_k; \xi_k)\|_2 \le \rho \|\nabla f_{\mathcal{S}_k}(w_k; \xi_k)\|_2.$$

Set $w_{k+1} \leftarrow w_k + \alpha_k s_k$, where $\alpha_k \in \{\gamma^0, \gamma^1, \gamma^2, \dots\}$ is the largest element with

$$f_{\mathcal{S}_k}(w_{k+1}; \xi_k) \le f_{\mathcal{S}_k}(w_k; \xi_k) + \eta \alpha_k \nabla f_{\mathcal{S}_k}(w_k; \xi_k)^T s_k. \quad (6.6)$$

end for

Stochastic Quasi-Newton Methods

- Approximate Hessian using only first-order methods
- Problems
  - Hessian approximations can be dense, even when Hessian is sparse
  - Limited memory scheme only allows provably linear convergence

$$s_k := w_{k+1} - w_k \quad \text{and} \quad v_k := \nabla F(w_{k+1}) - \nabla F(w_k),$$

$$H_{k+1} \leftarrow \left(I - \frac{v_k s_k^T}{s_k^T v_k}\right)^T H_k \left(I - \frac{v_k s_k^T}{s_k^T v_k}\right) + \frac{s_k s_k^T}{s_k^T v_k}$$

Gauss-Newton Methods

- Minimize second-order Taylor series expansion

$$G_{\mathcal{S}_k^H}(w_k;\xi_k^H) = \frac{1}{|\mathcal{S}_k^H|}\sum_{i\in\mathcal{S}_k^H} J_h(w_k;\xi_{k,i})^T H_\ell(w_k;\xi_{k,i})\, J_h(w_k;\xi_{k,i})$$

Natural Gradient Methods

- Invariant to all invertible transformations
- Gradient descent over prediction functions

$$w_{k+1} = \arg\min_{w \in \mathcal{W}} \; F(w) \quad \text{s.t.} \quad \tfrac{1}{2}(w - w_k)^T G(w_k)\,(w - w_k) \le \eta_k^2 \,.$$

$$w_{k+1} = \arg\min_{w \in \mathcal{W}} \; \nabla F(w_k)^T (w - w_k) + \frac{1}{2\alpha_k}(w - w_k)^T G(w_k)(w - w_k)$$

$$w_{k+1} = w_k - \alpha_k G^{-1}(w_k) \nabla F(w_k)$$

$$G(w) := \mathbb{E}_{h_w}\left[\frac{\partial^2 \log(h_w(x))}{\partial w^2}\right] = \mathbb{E}_{h_w}\left[\left(\frac{\partial \log(h_w(x))}{\partial w}\right)\left(\frac{\partial \log(h_w(x))}{\partial w}\right)^T\right]$$

Diagonal Scaling Methods

- Rescale search direction using diagonal transformation
- Examples
  - RMSProp
  - AdaGrad
- Structural Methods
  - Batch Normalization