Performance Limits of Single-Agent and Multi-Agent Sub-Gradient Stochastic Learning

Bicheng Ying and Ali H. Sayed

Electrical Engineering Department, University of California, Los Angeles, CA 90095

Emails: {ybc, sayed}@ucla.edu

Abstract

This work examines the performance of stochastic sub-gradient learning strategies, for both cases of stand-alone and networked agents, under weaker conditions than usually considered in the literature. It is shown that these conditions are automatically satisfied by several important cases of interest, including support-vector machines and sparsity-inducing learning solutions. The analysis establishes that sub-gradient strategies can attain exponential convergence rates, as opposed to sub-linear rates, and that they can approach the optimal solution within $O(\mu)$, for sufficiently small step-sizes, μ. A realizable exponential-weighting procedure is proposed to smooth the intermediate iterates and to guarantee these desirable performance properties.

Introduction

The minimization of *non-differentiable* convex cost functions is a critical step in the solution of many important design problems [1–3], including the design of sparse-aware (LASSO) solutions [4,5], support-vector machine (SVM) learners [6–10], or total-variation based image denoising solutions [11, 12]. The sub-gradient technique is a popular choice for minimizing such non-differentiable costs; it is closely related to the traditional gradient-descent method where the actual gradient vector is replaced by a sub-gradient at points of non-differentiability. It is one of the simplest methods in current practice but is known to suffer from slow convergence. In particular, it is shown in [3] that, for convex cost functions, the optimal convergence rate that can be delivered by sub-gradient methods in *deterministic* optimization problems cannot be faster than the $O(1/\sqrt{i})$, where i is the iteration index.

However, the results in subsequent sections will show that when used in the context of *stochastic* optimization, sub-gradient descent algorithms turn out to have superior performance than suggested by traditional analyses in the deterministic context. In particular, under constant step-size adaptation, these algorithms will be shown to converge at the faster exponential rate of $O(\alpha^i)$ for some $\alpha \in (0,1)$ when the cost function is strongly-convex. This rate is much faster than the $O(1/i)$ rate that would be observed under a diminishing step-size implementation for strongly-convex costs. We will clarify these favorable properties for both cases of stand-alone agents and networked agents [13–16].

Problem Formulation: Single Agent Case

We consider the problem of minimizing a risk function, $J(w): \mathbb{R}^M \rightarrow \mathbb{R}$, which is assumed to be expressed as the expected value of some loss function, $Q(w;\boldsymbol{x})$, namely,

$$w^\star \;\triangleq\; \arg\min_w J(w) \;\triangleq\; \arg\min_w \mathbb{E}_x\, Q(w;\boldsymbol{x}) \qquad (1)$$

where $w^\star$ denotes the minimizer. We first denote the sub-gradient of $J(w)$ at any arbitrary point w_0 by $g(w_0)$, and defined it as any vector $g \in \mathbb{R}^M$ that satisfies:

$$J(w) \;\geq\; J(w_0) + g^\mathsf{T}(w_0)(w - w_0), \quad \forall w \qquad (2)$$

In the context of adaptation and learning, we do not know the exact form of $J(w)$ because the distribution of the data is not known to enable computation of $\mathbb{E}_x Q(w;\boldsymbol{x})$. As such, true sub-gradient vectors for $J(w)$ cannot be determined and they will need to be replaced by stochastic approximations evaluated from streaming data. We employ the following stochastic iteration [1, 3, 24, 25]:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \mu\,\widehat{g}(\boldsymbol{w}_{i-1}) \qquad (3)$$

where the successive iterates, $\{\boldsymbol{w}_i\}$, are now random variables (denoted in boldface) and $\widehat{g}(\cdot)$ represents an approximate sub-gradient vector at location $\boldsymbol{w}_{i-1}$ estimated from data available at time i. The difference between an actual sub-gradient vector and its approximation is referred to as *gradient noise* and is denoted by

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \;\triangleq\; \widehat{g}(\boldsymbol{w}_{i-1}) - g(\boldsymbol{w}_{i-1}) \qquad (4)$$

Modeling Conditions and Analysis

Assumption 1 (CONDITIONS ON GRADIENT NOISE) *The first and second-order conditional moments of the gradient noise process satisfy the following conditions:*

$$\mathbb{E}[\,s_i(\boldsymbol{w}_{i-1})\,|\,\boldsymbol{\mathcal{F}}_{i-1}] \;=\; 0 \qquad (5)$$

$$\mathbb{E}[\,\|s_i(\boldsymbol{w}_{i-1})\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}] \;\leq\; \beta^2\|w^\star - \boldsymbol{w}_{i-1}\|^2 + \sigma^2 \qquad (6)$$

for some constants $\beta^2 \geq 0$ and $\sigma^2 \geq 0$, and where $\boldsymbol{\mathcal{F}}_{i-1}$ denotes the filtration corresponding to all past iterates (essentially, the conditioning in (5)–(6) is relative to the previous iterates). ■

The second condition ensures that $w^\star$ is unique so that the optimization problem is well-defined, and the third condition is more relaxed than what is traditionally imposed in the literature.

Assumption 2 (STRONGLY-CONVEX RISK FUNCTION) *The risk function is assumed to be $\eta-$strongly-convex, i.e.,*

$$J(\theta w_1 + (1-\theta)w_2) \;\leq\; \theta J(w_1) + (1-\theta)J(w_2) - \frac{\eta}{2}\theta(1-\theta)\|w_1 - w_2\|^2 \qquad (7)$$

for any $\theta \in [0,1]$, w_1, and w_2, and where $\eta > 0$ ■

Assumption 3 (SUB-GRADIENT IS AFFINE-LIPSCHITZ) *It is assumed that the sub-gradient of the risk function, $J(w)$, is affine Lipschitz, i.e. there exist constants $c \geq 0$ and $d \geq 0$ such that*

$$\|g(w_1) - g(w_2)\| \;\leq\; c\|w_1 - w_2\| + d, \quad \forall w_1, w_2 \qquad (8)$$

and for any choice $g(\cdot) \in \partial J(\cdot)$, where $\partial J(w)$ represent sub-differentials, i.e., the set of all valid sub-gradients at w. ■

In preparation for the analysis, we first conclude from (8) that:

$$\|g(w_1) - g(w_2)\|^2 \leq e^2\|w_1 - w_2\|^2 + f^2 \;\; \forall w_1, w_2,\; g \in \partial J \qquad (13)$$

where

$$e^2 \;\triangleq\; c^2 + \frac{2cd}{R} \geq 0, \quad f^2 \;\triangleq\; d^2 + 2cdR \geq 0 \qquad (14)$$

and the constant R is any positive number that we are free to choose.

Theorem 1 (SINGLE AGENT PERFORMANCE) *Consider using the stochastic sub-gradient algorithm (3) to seek the unique minimizer, $w^\star$, of problem (1), where the risk function satisfies Assumptions 1–3. If the step-size parameter is sufficiently small, then it holds that*

$$\lim_{i\to\infty} \mathbb{E}\, J(\boldsymbol{w}_i^{\text{best}}) - J(w^\star) \;\leq\; \mu(f^2 + \sigma^2)/2 \qquad (16)$$

Moreover, the convergence of $\mathbb{E}\, J(\boldsymbol{w}_i^{\text{best}})$ towards $J(w^\star)$ occurs at an exponential rate, $O(\alpha^i)$, where

$$\alpha \;\triangleq\; 1 - \mu\eta + \mu^2(e^2 + \beta^2) = 1 - O(\mu) \qquad (17)$$

Suppose we choose a parameter κ that satisfies $\alpha \leq \kappa < 1$. Next, we introduce the convex-combination coefficients:

$$r_L(j) \;\triangleq\; \frac{\kappa^{L-j}}{S_L}, \quad j = 0, 1, \ldots, L, \text{ where } S_L \;\triangleq\; \sum_{j=0}^{L} \kappa^{L-j} \qquad (18)$$

Using these coefficients, we define the weighted iterate

$$\bar{\boldsymbol{w}}_L \;\triangleq\; \textstyle\sum_{j=0}^{L} r_L(j)\boldsymbol{w}_j \qquad (19)$$

Under the same conditions as in Theorem 1, it holds that

$$\lim_{L\to\infty} \mathbb{E}\, J(\bar{\boldsymbol{w}}_L) - J(w^\star) \leq \mu(f^2 + \sigma^2)/2 \qquad (22)$$

The convergence of $\mathbb{E}\, J(\bar{\boldsymbol{w}}_L)$ towards $J(w^\star)$ continues to occur at an exponential rate.

Problem Formulation: Multi-Agent Case

We now extend the previous results to multi-agent networks where a collection of agents cooperate with each other to seek the minimizer of an aggregate cost of the form:

$$\min_w \sum_{k=1}^{N} J_k(w), \quad \text{where } J_k(w) \;\triangleq\; \mathbb{E}_{\boldsymbol{x}_k} Q_k(w; \boldsymbol{x}_k) \qquad (23)$$

We consider the following diffusion strategy in its adapt-then-combine (ATC) form:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\,\widehat{g}_k(\boldsymbol{w}_{k,i-1}) \qquad (24)$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\boldsymbol{\psi}_{\ell,i} \qquad (25)$$

Theorem 2 (NETWORK PERFORMANCE) *Consider using the stochastic sub-gradient diffusion algorithm (24)–(25) to seek the unique minimizer, $w^\star$, of problem (23), where the risk functions, $J_k(w)$, satisfy Assumptions 1–3 with parameters $\{\eta_k, \beta_k^2, \sigma_k^2, e_k^2, f_k^2\}$. Assume the step-size parameter is sufficiently small. It holds that*

$$\lim_{i\to\infty} \mathbb{E}\left(\sum_{k=1}^{N} p_k J_k(\boldsymbol{w}_{k,i}^{\text{best}}) - \sum_{k=1}^{N} p_k J_k(w^\star)\right) \leq \frac{\mu}{2}\sum_{k=1}^{N}\left(p_k f_k^2 + p_k^2\sigma_k^2 + 2p_k f_k h\right) \;=\; O(\mu) \qquad (27)$$

for some finite constant h. Moreover, the convergence occurs at an exponential rate, $O(\alpha_q^i)$, where

$$\alpha_q \;\triangleq\; \max_k \left\{1 - \mu\eta_k + \mu^2 e_k^2 + \mu^2\beta_k^2 p_k + \mu^2 h\frac{e_k^2}{f_k}\right\} \;=\; 1 - O(\mu) \qquad (28)$$

Application over SVM problem

The two-class SVM formulation deals with the problem of determining a separating hyperplane, $w \in \mathbb{R}^M$, in order to classify feature vectors, denoted by $\boldsymbol{h} \in \mathbb{R}^M$, into one of two classes: $\boldsymbol{\gamma} = +1$ or $\boldsymbol{\gamma} = -1$. The regularized SVM risk function is strongly-convex and of the form:

$$J^{\text{svm}}(w) \;\triangleq\; \frac{\rho}{2}\|w\|^2 + \mathbb{E}\left(\max\left\{0, 1 - \boldsymbol{\gamma}\boldsymbol{h}^\mathsf{T} w\right\}\right) \qquad (10)$$

We compare the performance of the stochastic sub-gradient SVM implementation against LIBSVM (a popular SVM solver that uses quadratic programming on dual problem) [27]. The test data is obtained from the LIBSVM website[1] and also from the UCI dataset[2]. We first use the Adult dataset after preprocessing [28] with 11,220 training data and 21,341 testing data in 123 feature dimensions.