Generalization Properties and Implicit Regularization for Multiple Passes SGM

Junhong Lin[*] JHLIN5@HOTMAIL.COM
Raffaello Camoriano[†,*,‡] RAFFAELLO.CAMORIANO@IIT.IT
Lorenzo Rosasco[*,‡] LROSASCO@MIT.EDU

[*]LCSL, Massachusetts Institute of Technology and Istituto Italiano di Tecnologia, Cambridge, MA 02139, USA
[‡]DIBRIS, Università degli Studi di Genova, Via Dodecaneso 35, Genova, Italy
[†]iCub Facility, Istituto Italiano di Tecnologia, Via Morego 30, Genova, Italy

Abstract

We study the generalization properties of stochastic gradient methods for learning with convex loss functions and linearly parameterized functions. We show that, in the absence of penalizations or constraints, the stability and approximation properties of the algorithm can be controlled by tuning either the step-size or the number of passes over the data. In this view, these parameters can be seen to control a form of implicit regularization. Numerical results complement the theoretical findings.

1. Introduction

The stochastic gradient method (SGM), often called stochastic gradient descent, has become an algorithm of choice in machine learning, because of its simplicity and small computational cost especially when dealing with big data sets (Bousquet & Bottou, 2008).
Despite its widespread use, the generalization properties of the variants of SGM used in practice are relatively little understood. Most previous works consider generalization properties of SGM with only one pass over the data, see e.g. (Nemirovski et al., 2009) or (Orabona, 2014) and references therein, while in practice multiple passes are usually considered. The effect of multiple passes has been studied extensively for the optimization of an empirical objective (Boyd & Mutapcic, 2007), but the role for generalization is less clear. In practice, early-stopping of the number of iterations, for example monitoring a hold-out set error, is a strategy often used to regularize. Moreover, the step-size is typically tuned to obtain the best results. The study in this paper is a step towards grounding theoretically these commonly used heuristics.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

Our starting points are a few recent works considering the generalization properties of different variants of SGM. One first series of results focus on least squares, either with one (Ying & Pontil, 2008; Tarres & Yao, 2014; Dieuleveut & Bach, 2014), or multiple (deterministic) passes over the data (Rosasco & Villa, 2015). In the former case it is shown that, in general, if only one pass over the data is considered, then the step-size needs to be tuned to ensure optimal results. In (Rosasco & Villa, 2015) it is shown that a universal step-size choice can be taken, if multiple passes are considered. In this case, it is the stopping time that needs to be tuned.

In this paper, we are interested in general, possibly non smooth, convex loss functions. The analysis for least squares heavily exploits properties of the loss and does not generalize to this broader setting. Here, our starting points are the results in (Lin et al., 2016; Hardt et al., 2016; Orabona, 2014) considering convex loss functions. In (Lin et al., 2016), early stopping of a (kernelized) batch subgradient method is analyzed, whereas in (Hardt et al., 2016) the stability properties of SGM for smooth loss functions are considered in a general stochastic optimization setting and certain convergence results are derived. In (Orabona, 2014), a more complex variant of SGM is analyzed and shown to achieve optimal rates.

Since we are interested in analyzing regularization and generalization properties of SGM, in this paper we consider a general non-parametric setting. In this latter setting, the effects of regularization are typically more evident since it can directly affect the convergence rates. In this context, the difficulty of a problem is characterized by an assumption on the approximation error. Under this condition, the need for regularization becomes clear. Indeed, in the absence of other constraints, the good performance of the algorithm relies on a bias-variance trade-off that can be controlled by suitably choosing the step-size and/or the number of passes. These latter parameters can be seen to act as regularization parameters.

Here, we refer to the regularization as 'implicit', in the sense that it is achieved neither by penalization nor by adding explicit constraints. The two main variants of the algorithm are the same as in least squares: one pass over the data with tuned step-size, or, fixed step-size choice and number of passes appropriately tuned. While in principle optimal parameter tuning requires explicitly solving a bias-variance trade-off, in practice adaptive choices can be implemented by cross-validation. In this case, both algorithm variants achieve optimal results, but different computations are entailed. In the first case, multiple single pass SGM need to be considered with different step-sizes, whereas in the second case, early stopping is used. Experimental results, complementing the theoretical analysis, are given and provide further insights on the properties of the algorithms.

The rest of the paper is organized as follows. In Section 2, we describe the supervised learning setting and the algorithm, and in Section 3, we state and discuss our main results. The proofs are postponed to the supplementary material. In Section 4, we present some numerical experiments on real datasets.

Notation. For notational simplicity, $[m]$ denotes $\{1, 2, \cdots, m\}$ for any $m \in \mathbb{N}$. The notation $a_k \lesssim b_k$ means that there exists a universal constant $C > 0$ such that $a_k \leq C b_k$ for all $k \in \mathbb{N}$. Denote by $\lceil a \rceil$ the smallest integer greater than a for any given $a \in \mathbb{R}$.

2. Learning with SGM

In this section, we introduce the supervised learning problem and the SGM algorithm.

Learning Setting. Let X be a probability space and Y be a subset of $\mathbb{R}$. Let ρ be a probability measure on $Z = X \times Y$. Given a measurable loss function $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, the associated expected risk $\mathcal{E} = \mathcal{E}^V$ is defined as

$$\mathcal{E}(f) = \int_Z V(y, f(x)) d\rho.$$

The distribution ρ is assumed to be fixed, but unknown, and the goal is to find a function minimizing the expected risk given a sample $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^m$ of size $m \in \mathbb{N}$ independently drawn according to ρ. Many classical examples of learning algorithms are based on empirical risk minimization, that is replacing the expected risk with the empirical risk $\mathcal{E}_\mathbf{z} = \mathcal{E}_\mathbf{z}^V$ defined as

$$\mathcal{E}_\mathbf{z}(f) = \frac{1}{m} \sum_{j=1}^m V(y_j, f(x_j)).$$

In this paper, we consider spaces of functions which are linearly parameterized. Consider a possibly non-linear data representation/feature map $\Phi : X \to \mathcal{F}$, mapping the data space in $\mathbb{R}^p$, $p \leq \infty$, or more generally in a (real separable) Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. Then, for $w \in \mathcal{F}$ we consider functions of the form

$$f_w(x) = \langle w, \Phi(x) \rangle, \quad \forall x \in X. \tag{1}$$

Examples of the above setting include the case where we consider infinite dictionaries, $\phi_j : X \to \mathbb{R}$, $j = 1, \ldots$, so that $\Phi(x) = (\phi_j(x))_{j=1}^\infty$, for all $x \in X$, $\mathcal{F} = \ell_2$ and (1) corresponds to $f_w = \sum_{j=1}^p w^j \phi_j$. Also, this setting includes, and indeed is equivalent to considering, functions defined by a positive definite kernel $K : X \times X \to \mathbb{R}$, in which case $\Phi(x) = K(x, \cdot)$, for all $x \in X$, $\mathcal{F} = \mathcal{H}_K$ the reproducing kernel Hilbert space associated with K, and (1) corresponds to the reproducing property

$$f_w(x) = \langle w, K(x, \cdot) \rangle, \forall x \in X. \tag{2}$$

In the following, we assume the feature map to be measurable and define expected and empirical risks over functions of the form (1). For notational simplicity, we write $\mathcal{E}(f_w)$ as $\mathcal{E}(w)$, and $\mathcal{E}_\mathbf{z}(f_w)$ as $\mathcal{E}_\mathbf{z}(w)$.

Stochastic Gradient Method. For any fixed $y \in Y$, assume the univariate function $V(y, \cdot)$ on $\mathbb{R}$ to be convex, hence its left-hand derivative $V'_-(y, a)$ exists at every $a \in \mathbb{R}$ and is non-decreasing.

Algorithm 1. *Given a sample* $\mathbf{z}$*, the stochastic gradient method (SGM) is defined by* $w_1 = 0$ *and*

$$w_{t+1} = w_t - \eta_t V'_-(y_{j_t}, \langle w_t, \Phi(x_{j_t}) \rangle) \Phi(x_{j_t}), \; t = 1, \ldots, T, \tag{3}$$

for a non-increasing sequence of step-sizes $\{\eta_t > 0\}_{t \in \mathbb{N}}$ *and a stopping rule* $T \in \mathbb{N}$*. Here,* $j_1, j_2, \cdots, j_T$ *are independent and identically distributed (i.i.d.) random variables*[1] *from the uniform distribution on* $[m]$*. The (weighted) averaged iterates are defined by*

$$\overline{w}_t = \sum_{k=1}^t \eta_k w_k / a_t, \quad a_t = \sum_{k=1}^t \eta_k, \quad t = 1, \ldots, T.$$

Note that T may be greater than m, indicating that we can use the sample more than once. We shall write $J(t)$ to mean $\{j_1, j_2, \cdots, j_t\}$, which will be also abbreviated as J when there is no confusion.

The main purpose of the paper is to estimate the expected excess risk of the last iterate

$$\mathbb{E}_{\mathbf{z}, J}[\mathcal{E}(w_T) - \inf_{w \in \mathcal{F}} \mathcal{E}(w)],$$

or similarly the expected excess risk of the averaged iterate $\overline{w}_T$, and study how different parameter settings in (1) affect

[1] More precisely, $j_1, j_2, \cdots, j_T$ are conditionally independent given any $\mathbf{z}$.

the estimates. Here, the expectation $\mathbb{E}_{\mathbf{z},J}$ stands for taking the expectation with respect to J (given any $\mathbf{z}$) first, and then the expectation with respect to $\mathbf{z}$.

3. Implicit Regularization for SGM

In this section, we present and discuss our main results. We begin in Subsection 3.1 with a universal convergence result and then provide finite sample bounds for smooth loss functions in Subsection 3.2, and for non-smooth functions in Subsection 3.3. As corollaries of these results we derive different implicit regularization strategies for SGM.

3.1. Convergence

We begin presenting a convergence result, involving conditions on both the step-sizes and the number of iterations. We need some basic assumptions.

Assumption 1. *There holds*

$$\kappa = \sup_{x\in X} \sqrt{\langle \Phi(x), \Phi(x)\rangle} < \infty. \tag{4}$$

Furthermore, the loss function is convex with respect to its second entry, and $|V|_0 := \sup_{y\in Y} V(y,0) < \infty$. *Moreover, its left-hand derivative* $V'_-(y,\cdot)$ *is bounded:*

$$\left|V'_-(y,a)\right| \leq a_0, \qquad \forall a \in \mathbb{R}, y \in Y. \tag{5}$$

The above conditions are common in statistical learning theory (Steinwart & Christmann, 2008; Cucker & Zhou, 2007). For example, they are satisfied for the hinge loss $V(y,a) = |1-ya|_+ = \max\{0, 1-ya\}$ or the logistic loss $V(y,a) = \log(1+\mathrm{e}^{-ya})$ for all $a \in \mathbb{R}$, if X is compact and $\Phi(x)$ is continuous.

The bounded derivative condition (5) is implied by the requirement on the loss function to be Lipschitz in its second entry, when Y is a bounded domain. Given these assumptions, the following result holds.

Theorem 1. *If Assumption 1 holds, then*

$$\lim_{m\to\infty} \mathbb{E}[\mathcal{E}(\overline{w}_{t^*(m)})] - \inf_{w\in\mathcal{F}} \mathcal{E}(w) = 0,$$

provided the sequence $\{\eta_k\}_k$ *and the stopping rule* $t^*(\cdot) : \mathbb{N} \to \mathbb{N}$ *satisfy*

(A) $\lim_{m\to\infty} \frac{\sum_{k=1}^{t^*(m)} \eta_k}{m} = 0,$

(B) and $\lim_{m\to\infty} \frac{1+\sum_{k=1}^{t^*(m)} \eta_k^2}{\sum_{k=1}^{t^*(m)} \eta_k} = 0.$

As seen from the proof in the appendix, Conditions (A) and (B) arise from the analysis of suitable sample, computational, and approximation errors. Condition (B) is similar to the one required by stochastic gradient methods (Bertsekas, 1999; Boyd et al., 2003; Boyd & Mutapcic, 2007). The difference is that here the limit is taken with respect to the number of points, but the number of passes on the data can be bigger than one.

Theorem 1 shows that in order to achieve consistency, the step-sizes and the running iterations need to be appropriately chosen. For instance, given m sample points for SGM with one pass[2], i.e., $t^*(m) = m$, possible choices for the step-sizes are $\{\eta_k = m^{-\alpha} : k \in [m]\}$ and $\{\eta_k = k^{-\alpha} : k \in [m]\}$ for some $\alpha \in (0,1)$. One can also fix the step-sizes *a priori*, and then run the algorithm with a suitable stopping rule $t^*(m)$.

These different parameter choices lead to different implicit regularization strategies as we discuss next.

3.2. Finite Sample Bounds for Smooth Loss Functions

In this subsection, we give explicit finite sample bounds for smooth loss functions, considering a suitable assumption on the approximation error.

Assumption 2. *The approximation error associated to the triplet* (ρ, V, Φ) *is defined by*

$$\mathcal{D}(\lambda) = \inf_{w\in\mathcal{F}} \left\{ \mathcal{E}(w) + \frac{\lambda}{2}\|w\|^2 \right\} - \inf_{w\in\mathcal{F}} \mathcal{E}(w), \quad \forall \lambda \geq 0. \tag{6}$$

We assume that for some $\beta \in (0,1]$ *and* $c_\beta > 0$, *the approximation error satisfies*

$$\mathcal{D}(\lambda) \leq c_\beta \lambda^\beta, \qquad \forall\, \lambda > 0. \tag{7}$$

Intuitively, Condition (7) quantifies how hard it is to achieve the infimum of the expected risk. In particular, it is satisfied with $\beta = 1$ when[3] $\exists w^* \in \mathcal{F}$ such that $\inf_{w\in\mathcal{F}} \mathcal{E}(w) = \mathcal{E}(w^*)$. More formally, the condition is related to classical terminologies in approximation theory, such as K-functionals and interpolation spaces (Steinwart & Christmann, 2008; Cucker & Zhou, 2007). The following remark is important for later discussions.

Remark 1 (SGM and Implicit Regularization)**.** *Assumption 2 is standard in statistical learning theory when analyzing Tikhonov regularization (Cucker & Zhou, 2007; Steinwart & Christmann, 2008). Besides, it has been shown that Tikhonov regularization can achieve best performance by choosing an appropriate penalty parameter which depends on the unknown parameter* β *(Cucker & Zhou, 2007; Steinwart & Christmann, 2008). In other words, in Tikhonov regularization, the penalty*

[2] We slightly abuse the term 'one pass', to mean m iterations.

[3] The existence of at least one minimizer in $\mathcal{F}$ is met for example when $\mathcal{F}$ is compact, or finite dimensional. In general, β does not necessarily have to be 1, since the hypothesis space may be chosen as a general infinite dimensional space, for example in non-parametric regression.

parameter plays a role of regularization. In this view, our coming results show that SGM can implicitly implement a form of Tikhonov regularization by controlling the step-size and/or the number of passes.

A further assumption relates to the smoothness of the loss, and is satisfied for example by the logistic loss.

Assumption 3. *For all $y \in Y$, $V(y,\cdot)$ is differentiable and $V'(y,\cdot)$ is Lipschitz continuous with a constant $L > 0$, i.e.*

$$|V'(y,b) - V'(y,a)| \leq L|b-a|, \quad \forall a, b \in \mathbb{R}.$$

The following result characterizes the excess risk of both the last and the average iterate for any fixed step-size and stopping time.

Theorem 2. *If Assumptions 1, 2 and 3 hold and $\eta_t \leq 2/(\kappa^2 L)$ for all $t \in \mathbb{N}$, then for all $t \in \mathbb{N}$,*

$$\mathbb{E}[\mathcal{E}(\overline{w}_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)] \lesssim \frac{\sum_{k=1}^t \eta_k}{m} + \frac{\sum_{k=1}^t \eta_k^2}{\sum_{k=1}^t \eta_k} + \left(\frac{1}{\sum_{k=1}^t \eta_k}\right)^\beta,$$

and

$$\mathbb{E}[\mathcal{E}(w_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)] \lesssim \frac{\sum_{k=1}^t \eta_k}{m} \sum_{k=1}^{t-1} \frac{\eta_k}{\eta_t(t-k)} + \left(\sum_{k=1}^{t-1} \frac{\eta_k^2}{\eta_t(t-k)} + \eta_t\right) + \frac{\left(\sum_{k=1}^t \eta_k\right)^{1-\beta}}{\eta_t t}.$$

The proof of the above result follows more or less directly from combining ideas and results in (Lin et al., 2016; Hardt et al., 2016) and is postponed to the appendix. The constants in the bounds are omitted, but given explicitly in the proof. While the error bound for the weighted average looks more concise than the one for the last iterate, interestingly, both error bounds lead to similar generalization properties.

The error bounds are composed of three terms related to sample error, computational error, and approximation error. Balancing these three error terms to achieve the minimum total error bound leads to optimal choices for the step-sizes $\{\eta_k\}$ and total number of iterations t^*. In other words, both the step-sizes $\{\eta_k\}$ and the number of iterations t^* can play the role of a regularization parameter. Using the above theorem, general results for step-size $\eta_k = \eta t^{-\theta}$ with some $\theta \in [0,1), \eta = \eta(m) > 0$ can be found in Proposition 3 from the appendix. Here, as corollaries we provide four different parameter choices to obtain the best bounds, corresponding to four different regularization strategies.

The first two corollaries correspond to fixing the step-sizes *a priori* and using the number of iterations as a regularization parameter. In the first result, the step-size is constant and depends on the number of sample points.

Corollary 1. *If Assumptions 1, 2 and 3 hold and $\eta_t = \eta_1/\sqrt{m}$ for all $t \in \mathbb{N}$ for some positive constant $\eta_1 \leq 2/(\kappa^2 L)$, then for all $t \in \mathbb{N}$, and $g_t = \overline{w}_t$ (or w_t),*

$$\mathbb{E}[\mathcal{E}(g_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)] \lesssim \frac{t\log t}{\sqrt{m^3}} + \frac{\log t}{\sqrt{m}} + \left(\frac{\sqrt{m}}{t}\right)^\beta. \quad (8)$$

In particular, if we choose $t^* = \lceil m^{\frac{\beta+3}{2(\beta+1)}} \rceil$,

$$\mathbb{E}[\mathcal{E}(g_{t^*}) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)] \lesssim m^{-\frac{\beta}{\beta+1}} \log m. \quad (9)$$

In the second result the step-sizes decay with the iterations.

Corollary 2. *If Assumptions 1, 2 and 3 hold and $\eta_t = \eta_1/\sqrt{t}$ for all $t \in \mathbb{N}$ with some positive constant $\eta_1 \leq 2/(\kappa^2 L)$, then for all $t \in \mathbb{N}$, and $g_t = \overline{w}_t$ (or w_t),*

$$\mathbb{E}[\mathcal{E}(g_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)] \lesssim \frac{\sqrt{t}\log t}{m} + \frac{\log t}{\sqrt{t}} + \frac{1}{t^{\beta/2}}. \quad (10)$$

Particularly, when $t^* = \lceil m^{\frac{2}{\beta+1}} \rceil$, we have (9).

In both the above corollaries the step-sizes are fixed *a priori*, and the number of iterations becomes the regularization parameter controlling the total error. Ignoring the logarithmic factor, the dominating terms in the bounds (8), (10) are the sample and approximation errors, corresponding to the first and third terms of RHS. Stopping too late may lead to a large sample error, while stopping too early may lead to a large approximation error. The ideal stopping time arises from a form of bias-variance trade-off and requires in general more than one pass over the data. Indeed, if we reformulate the results in terms of number of passes, we have that $\lceil m^{\frac{1-\beta}{2(1+\beta)}} \rceil$ passes are needed for the constant step-size $\{\eta_t = \eta_1/\sqrt{m}\}_t$, while $\lceil m^{\frac{1-\beta}{1+\beta}} \rceil$ passes are needed for the decaying step-size $\{\eta_t = \eta_1/\sqrt{t}\}_t$. These observations suggest in particular that while both step-size choices achieve the same bounds, the constant step-size can have a computational advantage since it requires less iterations.

Note that one pass over the data suffices only in the limit case when $\beta = 1$, while in general it will be suboptimal, at least if the step-size is fixed. In fact, Theorem 2 suggests that optimal results could be recovered if the step-size is suitably tuned. The next corollaries show that this is indeed the case. The first result corresponds to a suitably tuned constant step-size.

Corollary 3. *If Assumptions 1, 2 and 3 hold and* $\eta_t = \eta_1 m^{-\frac{\beta}{\beta+1}}$ *for all* $t \in \mathbb{N}$ *for some positive constant* $\eta_1 \leq 2/(\kappa^2 L)$*, then for all* $t \in \mathbb{N}$*, and* $g_t = \overline{w}_t$ *(or* w_t*),*

$$\begin{aligned}&\mathbb{E}[\mathcal{E}(g_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)]\\ \lesssim& m^{-\frac{\beta+2}{\beta+1}} t\log t + m^{-\frac{\beta}{\beta+1}}\log t + m^{\frac{\beta^2}{\beta+1}} t^{-\beta}.\end{aligned}$$

In particular, we have (9) for $t^* = m$*.*

The second result corresponds to tuning the decay rate for a decaying step-size.

Corollary 4. *If Assumptions 1, 2 and 3 hold and* $\eta_t = \eta_1 t^{-\frac{\beta}{\beta+1}}$ *for all* $t \in \mathbb{N}$ *for some positive constant* $\eta_1 \leq 2/(\kappa^2 L)$*, then for all* $t \in \mathbb{N}$*, and* $g_t = \overline{w}_t$ *(or* w_t*),*

$$\begin{aligned}&\mathbb{E}[\mathcal{E}(g_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)]\\ \lesssim& m^{-1} t^{\frac{1}{\beta+1}}\log t + t^{-\frac{\beta}{\beta+1}}\log t + t^{-\frac{\beta}{\beta+1}}.\end{aligned}$$

In particular, we have (9) for $t^* = m$*.*

The above two results confirm that good performances can be attained with only one pass over the data, provided the step-sizes are suitably chosen, that is using the step-size as a regularization parameter.

Remark 2. *If we further assume that* $\beta = 1$*, as often done in the literature, the convergence rates from Corollaries 1-4 are of order* $O(m^{-1/2})$*, which are the same as those in, e.g., (Shamir & Zhang, 2013).*

Finally, the following remark relates the above results to data-driven parameter tuning used in practice.

Remark 3 (Bias-Variance and Cross-Validation)**.** *The above results show how the number of iterations/passes controls a bias-variance trade-off, and in this sense acts as a regularization parameter. In practice, the approximation properties of the algorithm are unknown and the question arises of how the parameter can be chosen. As it turns out, cross-validation can be used to achieve adaptively the best rates, in the sense that the rate in (9) is achieved by cross-validation or more precisely by hold-out cross-validation. These results follow by an argument similar to that in Chapter 6 from (Steinwart & Christmann, 2008) and are omitted.*

3.3. Finite Sample Bounds for Non-smooth Loss Functions

Theorem 2 holds for smooth loss functions and it is natural to ask if a similar result holds for non-smooth losses such as the hinge loss. Indeed, analogous results hold, albeit current bounds are not as sharp.

Theorem 3. *If Assumptions 1 and 2 hold, then* $\forall t \in \mathbb{N}$*,*

$$\begin{aligned}&\mathbb{E}[\mathcal{E}(\overline{w}_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)]\\ \lesssim& \sqrt{\frac{\sum_{k=1}^t \eta_k}{m}} + \frac{\sum_{k=1}^t \eta_k^2}{\sum_{k=1}^t \eta_k} + \left(\frac{1}{\sum_{k=1}^t \eta_k}\right)^{\beta},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[\mathcal{E}(w_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)] \lesssim& \sqrt{\frac{\sum_{k=1}^t \eta_k}{m}} \sum_{k=1}^{t-1} \frac{\eta_k}{\eta_t (t-k)}\\ &+ \sum_{k=1}^{t-1} \frac{\eta_k^2}{\eta_t (t-k)} + \eta_t + \frac{\left(\sum_{k=1}^t \eta_k\right)^{1-\beta}}{\eta_t t}.\end{aligned}$$

The proof of the above theorem is based on ideas from (Lin et al., 2016), where tools from Rademacher complexity (Bartlett & Mendelson, 2003; Meir & Zhang, 2003) are employed. We postpone the proof in the appendix.

Using the above result with concrete step-sizes as those for smooth loss functions, we have the following explicit error bounds and corresponding stopping rules.

Corollary 5. *Under Assumptions 1 and 2, let* $\eta_t = 1/\sqrt{m}$ *for all* $t \in \mathbb{N}$*. Then for all* $t \in \mathbb{N}$*, and* $g_t = \overline{w}_t$ *(or* w_t*),*

$$\mathbb{E}[\mathcal{E}(g_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)] \lesssim \frac{\sqrt{t}\log t}{m^{3/4}} + \frac{\log t}{\sqrt{m}} + \left(\frac{\sqrt{m}}{t}\right)^{\beta}.$$

In particular, if we choose $t^* = \lceil m^{\frac{2\beta+3}{4\beta+2}} \rceil$*,*

$$\mathbb{E}[\mathcal{E}(g_{t^*}) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)] \lesssim m^{-\frac{\beta}{2\beta+1}} \log m. \qquad (11)$$

Corollary 6. *Under Assumptions 1 and 2, let* $\eta_t = 1/\sqrt{t}$ *for all* $t \in \mathbb{N}$*. Then for all* $t \in \mathbb{N}$*, and* $g_t = \overline{w}_t$ *(or* w_t*),*

$$\mathbb{E}[\mathcal{E}(g_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)] \lesssim \frac{t^{1/4}\log t}{\sqrt{m}} + \frac{\log t}{\sqrt{t}} + \frac{1}{t^{\beta/2}}.$$

In particular, if we choose $t^* = \lceil m^{\frac{2}{2\beta+1}} \rceil$*, there holds (11).*

From the above two corollaries, we see that the algorithm with constant step-size $1/\sqrt{m}$ can stop earlier than the one with decaying step-size $1/\sqrt{t}$ when $\beta \leq 1/2$, while they have the same convergence rate, since $m^{\frac{2\beta+3}{4\beta+2}}/m^{\frac{2}{2\beta+1}} = m^{\frac{2\beta-1}{4\beta+1}}$. Note that the bound in (11) is slightly worse than that in (9), see Section 3.4 for more discussion.

Similar to the smooth case, we also have the following results for SGM with one pass where regularization is realized by step-size.

Corollary 7. *Under Assumptions 1 and 2, let* $\eta_t = m^{-\frac{2\beta}{2\beta+1}}$ *for all* $t \in \mathbb{N}$*. Then for all* $t \in \mathbb{N}$*, and* $g_t = \overline{w}_t$ *(or*

w_t),

$$\mathbb{E}[\mathcal{E}(g_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)]$$
$$\lesssim m^{-\frac{4\beta+1}{4\beta+2}}\sqrt{t}\log t + m^{-\frac{2\beta}{2\beta+1}}\log t + m^{\frac{2\beta^2}{2\beta+1}}t^{-\beta}.$$

In particular, (11) holds for $t^* = m$.

Corollary 8. *Under Assumptions 1 and 2, let $\eta_t = t^{-\frac{2\beta}{2\beta+1}}$ for all $t \in \mathbb{N}$. Then for all $t \in \mathbb{N}$, and $g_t = \overline{w}_t$ (or w_t),*

$$\mathbb{E}[\mathcal{E}(g_t) - \inf_{w\in\mathcal{F}} \mathcal{E}(w)]$$
$$\lesssim m^{-\frac{1}{2}}t^{\frac{1}{4\beta+2}}\log t + t^{-\frac{\min(2\beta,1)}{2\beta+1}}\log t + t^{-\frac{\beta}{2\beta+1}}.$$

In particular, (11) holds for $t^* = m$.

3.4. Discussion and Proof Sketch

As mentioned in the introduction, the literature on theoretical properties of the iteration in Algorithm 1 is vast, both in learning theory and in optimization. A first line of works focuses on a single pass and convergence of the expected risk. Approaches in this sense include classical results in optimization (see (Nemirovski et al., 2009) and references therein), but also approaches based on so-called "online to batch" conversion (see (Orabona, 2014) and references therein). The latter are based on analyzing a sequential prediction setting and then on considering the averaged iterate to turn regret bounds in expected risk bounds. A second line of works focuses on multiple passes, but measures the quality of the corresponding iteration in terms of the minimization of the empirical risk. In this view, Algorithm 1 is seen as an instance of incremental methods for the minimization of objective functions that are sums of a finite, but possibly large, number of terms (Bertsekas, 2011). These latter works, while interesting in their own right, do not yield any direct information on the generalization properties of considering multiple passes.

Here, we follow the approach in (Bousquet & Bottou, 2008) advocating the combination of statistical and computational errors. The general proof strategy is to consider several intermediate steps to relate the expected risk of the empirical iteration to the minimal expected risk. The argument we sketch below is a simplified and less sharp version with respect to the one used in the actual proof, but it is easier to illustrate and still carries some important aspects which are useful for comparison with related results.

Consider an intermediate element $\tilde{w} \in \mathcal{F}$ and decompose the excess risk as

$$\mathbb{E}\mathcal{E}(w_t) - \inf_{w\in\mathcal{F}} \mathcal{E} =$$
$$\mathbb{E}(\mathcal{E}(w_t) - \mathcal{E}_{\mathbf{z}}(w_t)) + \mathbb{E}(\mathcal{E}_{\mathbf{z}}(w_t) - \mathcal{E}_{\mathbf{z}}(\tilde{w}))$$
$$+\mathbb{E}\mathcal{E}_{\mathbf{z}}(\tilde{w}) - \inf_{w\in\mathcal{F}} \mathcal{E}.$$

The first term on the right-hand side is the generalization error of the iterate. The second term can be seen as a computational error. To discuss the last term, it is useful to consider a few different choices for $\tilde{w}$. Assuming the empirical and expected risks to have minimizers $w^*_{\mathbf{z}}$ and w^*, a possibility is to set $\tilde{w} = w^*_{\mathbf{z}}$, this can be seen to be the choice made in (Hardt et al., 2016). In this case, it is immediate to see that the last term is negligible since,

$$\mathbb{E}\mathcal{E}_{\mathbf{z}}(\tilde{w}) = \mathbb{E}\min_{w\in\mathcal{F}} \mathcal{E}_{\mathbf{z}}(w) \leq \min_{w\in\mathcal{F}} \mathbb{E}\mathcal{E}_{\mathbf{z}}(w) = \min_{w\in\mathcal{F}} \mathcal{E}(w),$$

and hence,

$$\mathbb{E}\mathcal{E}_{\mathbf{z}}(\tilde{w}) - \min_{w\in\mathcal{F}} \mathcal{E} \leq 0.$$

On the other hand, in this case the computational error depends on the norm $\|w^*_{\mathbf{z}}\|$ which is in general hard to estimate. A more convenient choice is to set $\tilde{w} = w^*$. A reasoning similar to the one above shows that the last term is still negligible and the computational error can still be controlled depending on $\|w^*\|$. In a non-parametric setting, the existence of a minimizer is not ensured and corresponds to a limit case where there is small approximation error. Our approach is then to consider an *almost* minimizer of the expected risk with a prescribed accuracy. Following (Lin et al., 2016), we do this introducing Assumption (6) and choosing $\tilde{w}$ as the unique minimizer of $\mathcal{E} + \lambda\|\cdot\|^2$, $\lambda > 0$. Then the last term in the error decomposition can be upper bounded by the approximation error.

For the generalization error, the stability results from (Hardt et al., 2016) provide sharp estimates for smooth loss functions and in the 'capacity independent' limit, that is under no assumptions on the covering numbers of the considered function space. For this setting, the obtained bound is optimal in the sense that it matches the best available bound for Tikhonov regularization (Steinwart & Christmann, 2008; Cucker & Zhou, 2007). For the non-smooth case a standard argument based on Rademacher complexity can be used, and easily extended to be capacity dependent. However, the corresponding bound is not sharp and improvements are likely to hinge on deriving better norm estimates for the iterates. The question does not seem to be straightforward and is deferred to a future work.

The computational error for the averaged iterates can be controlled using classic arguments (Boyd & Mutapcic, 2007), whereas for the last iterate the arguments in (Lin et al., 2016; Shamir & Zhang, 2013) are needed. Finally, Theorems 2, 3 result from estimating and balancing the various error terms with respect to the choice of the step-size and number of passes.

We conclude this section with some perspective on the results in the paper. We note that since the primary goal of this study was to analyze the implicit regularization effect of step-size and number of passes, we have considered a very simple iteration. However, it would be very

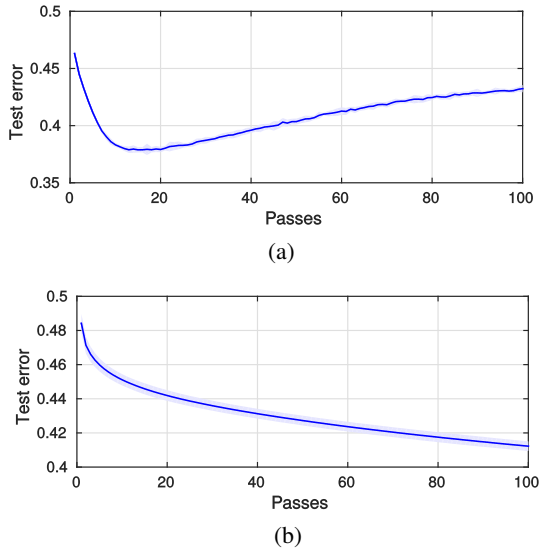

Figure 1. Test error for SIGM with fixed (a) and decaying (b) step-size with respect to the number of passes on *Adult* ($n = 1000$).

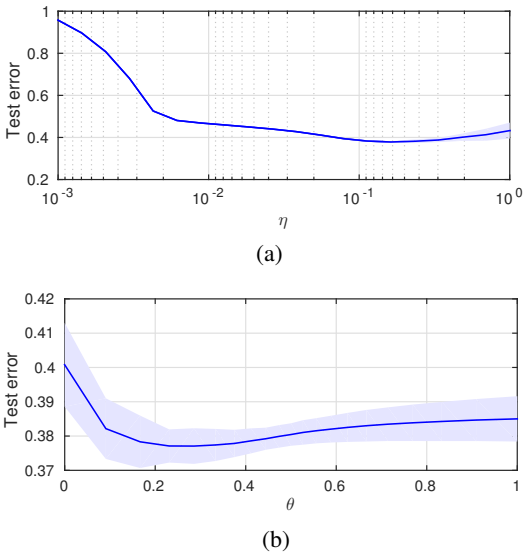

Figure 2. Test error for SGM with fixed (a) and decaying (b) step-size cross-validation on *Adult* ($n = 1000$).

interesting to consider more sophisticated, 'accelerated' iterations (Schmidt et al., 2013), and assess the potential advantages in terms of computational and generalization aspects. Similarly, we chose to keep the analysis in the paper relatively simple, but several improvements can be considered for example deriving high probability bounds and sharper error bounds under further assumptions. Some of these improvements are relatively straightforward, see e.g. (Lin et al., 2016), but others will require non-trivial extensions of results developed for Tikhonov regularization in the last few years. Finally, here we only referred to a simple cross-validation approach to parameter tuning, but it would clearly be very interesting to find ways to tune parameters online. A remarkable result in this direction is derived in (Orabona, 2014), where it is shown that, in the capacity independent setting, adaptive online parameter tuning is indeed possible.

4. Numerical Simulations

We carry out some numerical simulations to illustrate our results[4]. The experiments are executed 10 times each, on the benchmark datasets[5] reported in Table 1, in which the Gaussian kernel bandwidths σ used by SGM and SIGM[6] for each learning problem are also shown. Here, the loss function is the hinge loss[7]. The experimental platform is a server with 12 $\times$ Intel® Xeon® E5-2620 v2 (2.10GHz) CPUs and 132 GB of RAM. Some of the experimental results, as specified in the following, have been obtained by running the experiments on subsets of the data samples chosen uniformly at random. In order to apply hold-out cross-validation, the training set is split in two parts: one for empirical risk minimization and the other for validation error computation (80% - 20%, respectively). All the samples are randomly shuffled at each repetition.

Table 1. Benchmark datasets and Gaussian kernel width σ used in our experiments.

Dataset	n	n_{test}	d	σ
BreastCancer	400	169	30	0.4
Adult	32562	16282	123	4
Ijcnn1	49990	91701	22	0.6

4.1. Regularization in SGM and SIGM

In this subsection, we illustrate four concrete examples showing different regularization effects of the step-size in SGM and the number of passes in SIGM. In all these four examples, we consider the *Adult* dataset with sample size $n = 1000$.

In the first experiment, the SIGM step-size is fixed as $\eta = 1/\sqrt{n}$. The test error computed with respect to the

[4]Code: `lcsl.github.io/MultiplePassesSGM`

[5]Datasets: `archive.ics.uci.edu/ml` and `www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

[6]In what follows, we name one pass SGM and multiple passes SGM as SGM and SIGM, respectively.

[7]Experiments with the logistic loss have also been carried out, showing similar empirical results to those considering the hinge loss. The details are not included in this text due to space limit.

Table 2. Comparison of SGM and SIGM with cross-validation with decaying (D) and constant (C) step-sizes, in terms of computational time and accuracy. SGM performs cross-validation on 30 candidate step-sizes, while SIGM achieves implicit regularization via early stopping.

Dataset	*Algorithm*	*Step Size*	*Test Error (hinge loss)*	*Test Error (class. error)*	*Training Time (s)*
	SGM	C	0.127 ± 0.022	$3.1 \pm 1.1\%$	1.7 ± 0.2
BreastCancer	SGM	D	0.135 ± 0.024	$3.0 \pm 1.1\%$	1.4 ± 0.3
$n = 400$	$SIGM$	C	0.131 ± 0.023	$3.2 \pm 1.1\%$	1.4 ± 0.8
	$SIGM$	D	0.204 ± 0.017	$3.9 \pm 1.0\%$	1.8 ± 0.5
	$LIBSVM$			$2.8 \pm 1.3\%$	0.2 ± 0.0
	SGM	C	0.380 ± 0.003	$16.6 \pm 0.3\%$	5.7 ± 0.6
Adult	SGM	D	0.378 ± 0.002	$16.2 \pm 0.2\%$	5.4 ± 0.3
$n = 1000$	$SIGM$	C	0.383 ± 0.002	$16.1 \pm 0.0\%$	3.2 ± 0.4
	$SIGM$	D	0.450 ± 0.002	$23.6 \pm 0.0\%$	1.6 ± 0.2
	$LIBSVM$			$18.7 \pm 0.0\%$	5.8 ± 0.5
	SGM	C	0.342 ± 0.001	$15.2 \pm 0.8\%$	320.0 ± 3.3
Adult	SGM	D	0.340 ± 0.001	$15.1 \pm 0.7\%$	332.1 ± 3.3
$n = 32562$	$SIGM$	C	0.343 ± 0.001	$15.7 \pm 0.9\%$	366.2 ± 3.9
	$SIGM$	D	0.364 ± 0.001	$17.1 \pm 0.8\%$	442.4 ± 4.2
	$LIBSVM$			$15.3 \pm 0.7\%$	6938.7 ± 171.7
	SGM	C	0.199 ± 0.016	$8.4 \pm 0.8\%$	3.9 ± 0.3
Ijcnn1	SGM	D	0.199 ± 0.009	$9.1 \pm 0.1\%$	3.8 ± 0.3
$n = 1000$	$SIGM$	C	0.205 ± 0.010	$9.3 \pm 0.5\%$	1.7 ± 0.4
	$SIGM$	D	0.267 ± 0.006	$9.4 \pm 0.6\%$	2.2 ± 0.4
	$LIBSVM$			$7.1 \pm 0.7\%$	0.6 ± 0.1
	SGM	C	0.041 ± 0.002	$1.5 \pm 0.0\%$	564.9 ± 6.3
Ijcnn1	SGM	D	0.059 ± 0.000	$1.7 \pm 0.0\%$	578.9 ± 1.8
$n = 49990$	$SIGM$	C	0.098 ± 0.001	$4.7 \pm 0.1\%$	522.2 ± 20.7
	$SIGM$	D	0.183 ± 0.000	$9.5 \pm 0.0\%$	519.3 ± 25.8
	$LIBSVM$			$0.9 \pm 0.0\%$	770.4 ± 38.5

hinge loss at each pass is reported in Figure 1(a). Note that the minimum test error is reached for a number of passes smaller than 20, after which it significantly increases, a so-called overfitting regime. This result clearly illustrates the regularization effect of the number of passes. In the second experiment, we consider SIGM with decaying step-size ($\eta = 1/4$ and $\theta = 1/2$). As shown in Figure 1(b), overfitting is not observed in the first 100 passes. In this case, the convergence to the optimal solution appears slower than that in the fixed step-size case.

In the last two experiments, we consider SGM and show that the step-size plays the role of a regularization parameter. For the fixed step-size case, i.e., $\theta = 0$, we perform SGM with different $\eta \in (0, 1]$ (logarithmically scaled). We plot the errors in Figure 2(a), showing that a large step-size ($\eta = 1$) leads to overfitting, while a smaller one (e. g., $\eta = 10^{-3}$) is associated to oversmoothing. For the decaying step-size case, we fix $\eta_1 = 1/4$, and run SGM with different $\theta \in [0, 1]$. The errors are plotted in Figure 2(b), from which we see that the exponent θ has a regularization effect. In fact, a more 'aggressive' choice (e. g., $\theta = 0$, corresponding to a fixed step-size) leads to overfitting, while for a larger θ (e. g., $\theta = 1$) we observe oversmoothing.

4.2. Accuracy and Computational Time Comparison

In this subsection, we compare SGM with cross-validation and SIGM with benchmark algorithm LIBSVM (Chang & Lin, 2011), both in terms of accuracy and computational time. For SGM, with 30 parameter guesses, we use cross-validation to tune the step-size (either setting $\theta = 0$ while tuning η, or setting $\eta = 1/4$ while tuning θ). For SIGM, we use two kinds of step-size suggested by Section 3: $\eta = 1/\sqrt{m}$ and $\theta = 0$, or $\eta = 1/4$ and $\theta = 1/2$, using early stopping via cross-validation. The test errors with respect to the hinge loss, the test relative misclassification errors and the computational times are collected in Table 2.

We first start comparing accuracies. The results in Table 2 indicate that SGM with constant and decaying step-sizes and SIGM with fixed step-size reach comparable test errors, which are in line with the LIBSVM baseline. Observe that SIGM with decaying step-size attains consistently higher test errors, a phenomenon already illustrated in Section 4.1 in theory.

We now compare the computational times for cross-validation. We see from Table 2 that the training times of SIGM and SGM, either with constant or decaying step-sizes, are roughly the same. We also observe that SGM and SIGM are faster than LIBSVM on relatively large datasets (*Adult with n = 32562*, and *Ijcnn1 with n = 49990*). Moreover, for small datasets (*BreastCancer* with *n = 400*, *Adult* with *n = 1000*, and *Ijcnn1* with *n = 1000*), SGM and SIGM are comparable with or slightly slower than LIBSVM.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. L. R. acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC. The authors would like to thank Dr. Francesco Orabona for the fruitful discussions on this research topic, and Dr. Silvia Villa and the referees for their valuable comments.

References

Bartlett, Peter L and Mendelson, Shahar. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3: 463–482, 2003.

Bartlett, Peter L, Bousquet, Olivier, and Mendelson, Shahar. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

Bertsekas, Dimitri P. *Nonlinear Programming*. Athena

scientific, 1999.

Bertsekas, Dimitri P. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010:1–38, 2011.

Bousquet, Olivier and Bottou, Léon. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pp. 161–168, 2008.

Boyd, Stephen and Mutapcic, Almir. Stochastic subgradient methods. Notes for EE364b, Standford University, Winter 2007.

Boyd, Stephen, Xiao, Lin, and Mutapcic, Almir. Subgradient methods. Lecture notes of EE392o, Stanford University, Autumn Quarter 2003.

Chang, Chih-Chung and Lin, Chih-Jen. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

Cucker, Felipe and Zhou, Ding-Xuan. *Learning Theory: an Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007.

Dieuleveut, Aymeric and Bach, Francis. Non-parametric stochastic approximation with large step sizes. *arXiv preprint arXiv:1408.0361*, 2014.

Hardt, Moritz, Recht, Benjamin, and Singer, Yoram. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2016.

Lin, Junhong, Rosasco, Lorenzo, and Zhou, Ding-Xuan. Iterative regularization for learning with convex loss functions. *The Journal of Machine Learning Research, To appear*, 2016.

Meir, Ron and Zhang, Tong. Generalization error bounds for Bayesian mixture algorithms. *The Journal of Machine Learning Research*, 4:839–860, 2003.

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

Orabona, Francesco. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pp. 1116–1124, 2014.

Rosasco, Lorenzo and Villa, Silvia. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pp. 1621–1629, 2015.

Schmidt, Mark, Roux, Nicolas Le, and Bach, Francis. Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388*, 2013.

Shamir, Ohad and Zhang, Tong. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 71–79, 2013.

Steinwart, Ingo and Christmann, Andreas. *Support Vector Machines*. Springer Science Business Media, 2008.

Tarres, Pierre and Yao, Yuan. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

Ying, Yiming and Pontil, Massimiliano. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.