HAL
open science

Constant Step Size Stochastic Gradient Descent for Probabilistic Modeling

Dmitry Babichev, Francis Bach

To cite this version:

Dmitry Babichev, Francis Bach. Constant Step Size Stochastic Gradient Descent for Probabilistic Modeling. UAI 2018 - Conference on Uncertainty in Artificial Intelligence, Aug 2018, Monterey, United States. hal-01929810

HAL Id: hal-01929810

https://inria.hal.science/hal-01929810

Submitted on 21 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Constant Step Size Stochastic Gradient Descent for Probabilistic Modeling

Dmitry Babichev
INRIA - ENS
PSL Research University
Paris, France
dmitry.babichev@inria.fr

Francis Bach
INRIA - ENS
PSL Research University
Paris, France
francis.bach@inria.fr

Abstract

Stochastic gradient methods enable learning probabilistic models from large amounts of data. While large step-sizes (learning rates) have shown to be best for least-squares (e.g., Gaussian noise) once combined with parameter averaging, these are not leading to convergent algorithms in general. In this paper, we consider generalized linear models, that is, conditional models based on exponential families. We propose averaging moment parameters instead of natural parameters for constant-step-size stochastic gradient descent. For finite-dimensional models, we show that this can sometimes (and surprisingly) lead to better predictions than the best linear model. For infinite-dimensional models, we show that it always converges to optimal predictions, while averaging natural parameters never does. We illustrate our findings with simulations on synthetic data and classical benchmarks with many observations.

1 INTRODUCTION

Faced with large amounts of data, efficient parameter estimation remains one of the key bottlenecks in the application of probabilistic models. Once cast as an optimization problem, for example through the maximum likelihood principle, difficulties may arise from the size of the model, the number of observations, or the potential non-convexity of the objective functions, and often all three (Koller and Friedman, 2009; Murphy, 2012).

In this paper we focus primarily on situations where the number of observations is large; in this context, stochastic gradient descent (SGD) methods which look at one sample at a time are usually favored for their cheap iteration cost. However, finding the correct step-size (sometimes referred to as the learning rate) remains a practical and theoretical challenge, for probabilistic modeling but also in most other situations beyond maximum likelihood (Bottou et al., 2016).

In order to preserve convergence, the step size γ_n at the n-th iteration typically has to decay with the number of gradient steps (here equal to the number of data points which are processed), typically as C/n^α for $\alpha \in [1/2, 1]$ (see, e.g., Bach and Moulines, 2011; Bottou et al., 2016). However, these often leads to slow convergence and the choice of α and C is difficult in practice. More recently, constant step-sizes have been advocated for their fast convergence towards a neighborhood of the optimal solution (Bach and Moulines, 2013), while it is a standard practice in many areas (Goodfellow et al., 2016). However, it is not convergent in general and thus small step-sizes are still needed to converge to a decent estimator.

Constant step-sizes can however be made to converge in one situation. When the functions to optimize are quadratic, like for least-squares regression, using a constant step-size combined with an averaging of all estimators along the algorithm can be shown to converge to the global solution with the optimal convergence rates (Bach and Moulines, 2013; Dieuleveut and Bach, 2016).

The goal of this paper is to explore the possibility of such global convergence with a constant step-size in the context of probabilistic modeling with exponential families, e.g., for logistic regression or Poisson regression (McCullagh, 1984). This would lead to the possibility of using probabilistic models (thus with a principled quantification of uncertainty) with rapidly converging algorithms. Our main novel idea is to replace the averaging of the *natural* parameters of the exponential family by the averaging of the *moment* parameters, which can also be formulated as averaging *predictions* instead of *estimators*. For example, in the context of predicting bi-

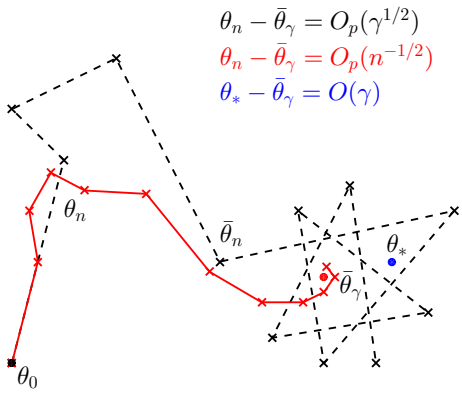

Figure 1: Convergence of iterates θ_n and averaged iterates $\bar{\theta}_n$ to the mean $\bar{\theta}_\gamma$ under the stationary distribution π_γ.

nary outcomes in $\{0, 1\}$ through a Bernoulli distribution, the moment parameter is the probability $p \in [0, 1]$ that the variable is equal to one, while the natural parameter is the "log odds ratio" $\log \frac{p}{1-p}$, which is unconstrained. This lack of constraint is often seen as a benefit for optimization; it turns out that for stochastic gradient methods, the moment parameter is better suited to averaging. Note that for least-squares, which corresponds to modeling with the Gaussian distribution with fixed variance, moment and natural parameters are equal, so it does not make a difference.

More precisely, our main contributions are:

- For generalized linear models, we propose in Section 4 averaging moment parameters instead of natural parameters for constant-step-size stochastic gradient descent.
- For finite-dimensional models, we show in Section 5 that this can sometimes (and surprisingly) lead to better predictions than the best linear model.
- For infinite-dimensional models, we show in Section 6 that it always converges to optimal predictions, while averaging natural parameters never does.
- We illustrate our findings in Section 7 with simulations on synthetic data and classical benchmarks with many observations.

2 CONSTANT STEP SIZE STOCHASTIC GRADIENT DESCENT

In this section, we present the main intuitions behind stochastic gradient descent (SGD) with constant step-size. For more details, see Dieuleveut et al. (2017). We consider a real-valued function F defined on the Euclidean space $\mathbb{R}^d$ (this can be generalized to any Hilbert space, as done in Section 6 when considering Gaussian processes and positive-definite kernels), and a sequence of random functions $(f_n)_{n \geqslant 1}$ which are independent and identically distributed and such that $\mathbb{E} f_n(\theta) = F(\theta)$ for all $\theta \in \mathbb{R}^d$. Typically, F will the expected negative log-likelihood on unseen data, while f_n will be the negative log-likelihood for a single observation. Since we require *independent* random functions, we assume that we make single pass over the data, and thus the number of iterations is equal to the number of observations.

Starting from an initial $\theta_0 \in \mathbb{R}^d$, then SGD will perform the following recursion, from $n = 1$ to the total number of observations:

$$\theta_n = \theta_{n-1} - \gamma_n \nabla f_n(\theta_{n-1}). \tag{1}$$

Since the functions f_n are independent, the iterates $(\theta_n)_n$ form a Markov chain. When the step-size γ_n is constant (equal to γ) and the functions f_n are identically distributed, the Markov chain is *homogeneous*. Thus, under additional assumptions (see, e.g., Dieuleveut et al., 2017; Meyn and Tweedie, 1993), it converges in distribution to a stationary distribution, which we refer to as π_γ. These additional assumptions include that γ is not too large (otherwise the algorithm diverges) and in the traditional analysis of step-sizes for gradient descent techniques, we analyze the situation of small γ's (and thus perform asymptotic expansions around $\gamma = 0$).

The distribution π_γ is in general not equal to a Dirac mass, and thus, constant-step-size SGD is *not* convergent. However, averaging along the path of the Markov chain has some interesting properties. Indeed, several versions of the "ergodic theorem" (see, e.g., Meyn and Tweedie, 1993) show that for functions g from $\mathbb{R}^d$ to any vector space, then the empirical average $\frac{1}{n} \sum_{i=1}^n g(\theta_i)$ converges in probability to the expectation $\int g(\theta) d\pi_\gamma(\theta)$ of g under the stationary distribution π_γ. This convergence can also be quantified by a central limit theorem with an error whichs tends to a normal distribution with variance proportional equal to a constant times $1/n$.

Thus, if denote $\bar{\theta}_n = \frac{1}{n+1} \sum_{i=0}^n \theta_i$, applying the previous result to the identity function g, we immediately obtain that $\bar{\theta}_n$ converges to $\bar{\theta}_\gamma = \int \theta d\pi_\gamma(\theta)$, with a squared error converging in $O(1/n)$. The key question is the relationship between $\bar{\theta}_\gamma$ and the global optimizer θ_* of F, as this characterizes the performance of the algorithm with an infinite number of observations.

By taking expectations in Eq. (1), and taking a limit with n tending to infinity we obtain that

$$\int \nabla F(\theta) d\pi_\gamma(\theta) = 0, \tag{2}$$

that is, under the stationary distribution π_γ, the average gradient is zero. When the gradient is a linear function (like for a quadratic objective F), this leads to $\nabla F(\int \theta d\pi_\gamma(\theta)) = \nabla F(\bar{\theta}_\gamma) = 0$, and thus $\bar{\theta}_\gamma$ is a stationary point of F (and hence the global minimizer if F is convex). However this is not true in general.

As shown by Dieuleveut et al. (2017), the deviation $\bar{\theta}_\gamma - \theta_*$ is of order γ, which is an improvement on the non-averaged recursion, which is at average distance $O(\gamma^{1/2})$ (see an illustration in Figure 1); thus, small or decaying step-sizes are needed. In this paper, we explore alternatives which are not averaging the estimators $\theta_1, \dots, \theta_n$, and rely instead on the specific structure of our cost functions, namely negative log-likelihoods.

3 WARM-UP: EXPONENTIAL FAMILIES

In order to highlight the benefits of averaging moment parameters, we first consider unconditional exponential families. We thus consider the standard exponential family $q(x|\theta) = h(x)\exp(\theta^\top T(x) - A(\theta))$, where $h(x)$ is the base measure, $T(x) \in \mathbb{R}^d$ is the sufficient statistics and A the log-partition function. The function A is always convex (see, e.g., Koller and Friedman, 2009; Murphy, 2012). Note that we do not assume that the data distribution $p(x)$ comes from this exponential family. The expected (with respect to the input distribution $p(x)$) negative log-likelihood is equal to

$$\begin{aligned} F(\theta) &= -\mathbb{E}_{p(x)} \log q(x|\theta) \\ &= A(\theta) - \theta^\top \mathbb{E}_{p(x)} T(x) - \mathbb{E}_{p(x)} \log h(x). \end{aligned}$$

It is known to be minimized by θ_* such that $\nabla A(\theta_*) = \mathbb{E}_{p(x)} T(x)$. Given i.i.d. data $(x_n)_{n \geqslant 1}$ sampled from $p(x)$, then the SGD recursion from Eq. (1) becomes:

$$\theta_n = \theta_{n-1} - \gamma\big[\nabla A(\theta_{n-1}) - T(x_n)\big],$$

while the stationarity equation in Eq. (2) becomes

$$\int \big[\nabla A(\theta) - \mathbb{E}_{p(x)} T(x)\big] d\pi_\gamma(\theta) = 0,$$

which leads to

$$\int \nabla A(\theta) d\pi_\gamma(\theta) = \mathbb{E}_{p(x)} T(x) = \nabla A(\theta_*).$$

Thus, averaging $\nabla A(\theta_n)$ will converge to $\nabla A(\theta_*)$, while averaging θ_n will *not* converge to θ_*. This simple observation is the basis of our work.

Note that in this context of unconditional models, a simpler estimator exists, that is, we can simply compute the empirical average $\frac{1}{n}\sum_{i=1}^n T(x_i)$ that will converge to $\nabla A(\theta_*)$. Nevertheless, this shows that averaging moment parameters $\nabla A(\theta)$ rather than natural parameters θ can bring convergence benefits. We now turn to conditional models, for which no closed-form solutions exist.

4 CONDITIONAL EXPONENTIAL FAMILIES

Now we consider the conditional exponential family $q(y|x,\theta) = h(y)\exp\big(y \cdot \eta_\theta(x) - a(\eta_\theta(x))\big)$. For simplicity we consider only one-dimensional families where $y \in \mathbb{R}$—but our framework would also extend to more complex models such as conditional random fields (Lafferty et al., 2001). We will also assume that $h(y) = 1$ for all y to avoid carrying constant terms in log-likelihoods. We consider functions of the form $\eta_\theta(x) = \theta^\top \Phi(x)$, which are linear in a feature vector $\Phi(x)$, where $\Phi : \mathcal{X} \to \mathbb{R}^d$ can be defined on an arbitrary input set $\mathcal{X}$. Calculating the negative log-likelihood, one obtains:

$$f_n(\theta) = -\log q(y_n|x_n\theta) = -y_n \Phi(x_n)^\top \theta + a\big(\Phi(x_n)\theta\big),$$

and, for any distribution $p(x,y)$, for which $p(y|x)$ may not be a member of the conditional exponential family,

$$\begin{aligned} F(\theta) &= \mathbb{E}_{p(x_n,y_n)} f_n(\theta) \\ &= \mathbb{E}_{p(x_n,y_n)}\Big[-y_n \Phi(x_n)^\top \theta + a\big(\Phi(x_n)\theta\big)\Big]. \end{aligned}$$

The goal of estimation in such generalized linear models is to find an unknown parameter θ given n observations $(x_i, y_i)_{i=1,\dots,n}$:

$$\theta_* = \arg\min_{\theta \in \mathbb{R}^d} F(\theta). \tag{3}$$

4.1 FROM ESTIMATORS TO PREDICTION FUNCTIONS

Another point of view is to consider that an estimator $\theta \in \mathbb{R}^d$ in fact defines a function $\eta : \mathcal{X} \to \mathbb{R}$, with value a natural parameter for the exponential family $q(y) = \exp(\eta y - a(\eta))$. This particular choice of function η_θ is linear in $\Phi(x)$, and we have, by decomposing the joint probability $p(x_n, y_n)$ in two (and dropping the dependence on n since we have assumed i.i.d. data):

$$\begin{aligned} F(\theta) &= \mathbb{E}_{p(x)}\Big(\mathbb{E}_{p(y|x)}\big[-y\Phi(x)^\top\theta + a(\Phi(x)^\top\theta)\big]\Big) \\ &= \mathbb{E}_{p(x)}\Big(-\mathbb{E}_{p(y|x)} y\Phi(x)^\top\theta + a(\Phi(x)^\top\theta)\Big) \\ &= \mathcal{F}(\eta_\theta), \end{aligned}$$

with $\mathcal{F}(\eta) = \mathbb{E}_{p(x)}\big(-\mathbb{E}_{p(y|x)} y \cdot \eta(x) + a(\eta(x))\big)$ is the performance measure defined for a *function* $\eta : \mathcal{X} \to \mathbb{R}$. By definition $F(\theta) = \mathcal{F}(\eta_\theta) = \mathcal{F}(\theta^\top \Phi(\cdot))$.

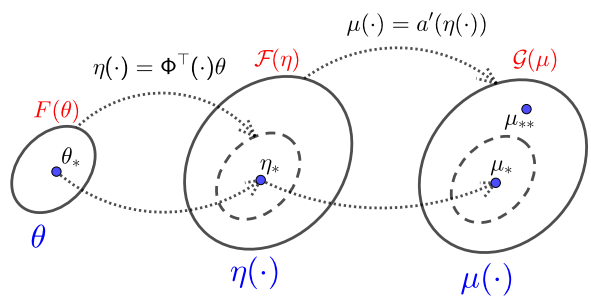

Figure 2: Graphical representation of reparametrization: firstly we expand the class of functions, replacing parameter θ with function $\eta(\cdot) = \Phi^\top(\cdot)\theta$ and then we do one more reparametrization: $\mu(\cdot) = a'(\eta(\cdot))$. Best linear prediction μ_* is constructed using θ_* and the global minimizer of $\mathcal{G}$ is μ_{**}. Model is well-specified if and only if $\mu_* = \mu_{**}$.

However, the global minimizer of $\mathcal{F}(\eta)$ over all functions $\eta : \mathcal{X} \to \mathbb{R}$ may not be attained at a linear function in $\Phi(x)$ (this can only be the case if the linear model is well-specified or if the feature vector $\Phi(x)$ is flexible enough). Indeed, the global minimizer of $\mathcal{F}$ is the function $\eta_{**} : x \mapsto (a')^{-1}(\mathbb{E}_{p(y|x)}y)$ (starting from $\mathcal{F}(\eta) = \int \big[a(\eta(x)) - \mathbb{E}_{p(x|y)}y \cdot \eta(x)\big]p(x)dx$ and writing down the Euler - Lagrange equation: $\frac{\partial \mathcal{F}}{\partial \eta} - \frac{d}{dx}\frac{\partial F}{\partial \eta'} = 0 \Leftrightarrow \big[a'(\eta) - \mathbb{E}_{p(x|y)}y\big]p(x) = 0$ and finally $\eta \mapsto (a')^{-1}(\mathbb{E}_{p(x|y)}y))$ and is typically not a linear function in $\Phi(x)$ (note here that $p(y|x)$ is the conditional data-generating distribution).

The function η corresponds to the *natural* parameter of the exponential family, and it is often more intuitive to consider the *moment* parameter, that is defining functions $\mu : \mathcal{X} \to \mathbb{R}$ that now correspond to moments of outputs y; we will refer to them as *prediction functions*. Going from natural to moment parameter is known to be done through the gradient of the log-partition function, and we thus consider for η a function from $\mathcal{X}$ to $\mathbb{R}$, $\mu(\cdot) = a'(\eta(\cdot))$, and this leads to the performance measure

$$\mathcal{G}(\mu) = \mathcal{F}((a')^{-1}(\mu(\cdot))).$$

Note now, that the global minimum of $\mathcal{G}$ is reached at

$$\mu_{**}(x) = \mathbb{E}_{p(y|x)}y.$$

We introduce also the prediction function $\mu_*(x)$ corresponding to the best η which is linear in $\Phi(x)$, that is:

$$\mu_*(x) = a'\big(\theta_*^\top \Phi(x)\big).$$

We say that the model is well-specified when $\mu_* = \mu_{**}$, and for these models, $\inf_\theta F(\theta) = \inf_\mu \mathcal{G}(\mu)$. However, in general, we only have $\inf_\theta F(\theta) > \inf_\mu \mathcal{G}(\mu)$ and (very often) the inequality is strict (see examples in our simulations).

To make the further developments more concrete, we now present two classical examples: logistic regression and Poisson regression.

Logistic regression. The special case of conditional family is logistic regression, where $y \in \{0, 1\}$, $a(t) = \log(1 + e^{-t})$ and $a'(t) = \sigma(t) = \frac{1}{1+e^{-t}}$ is the sigmoid function and the probability mass function is given by $p(y|\eta) = \exp(\eta y - \log(1 + e^\eta))$.

Poisson regression. One more special case is Poisson regression with $y \in \mathbb{N}$, $a(t) = \exp(t)$ and the response variable y has a Poisson distribution. The probability mass function is given by $p(y|\eta) = \exp(\eta y - e^\eta - \log(y!))$. Poisson regression may be appropriate when the dependent variable is a count, for example in genomics, network packet analysis, crime rate analysis, fluorescence microscopy, etc. (Hilbe, 2011).

4.2 AVERAGING PREDICTIONS

Recall from Section 2 that π_γ is the stationary distribution of θ. Taking expectation of both parts of Eq. (1), we get, by using the fact that π_γ is the limiting distribution of θ_n and θ_{n-1}:

$$\begin{aligned} & \mathbb{E}_{\pi_\gamma(\theta_n)}\theta_n \\ = \; & \mathbb{E}_{\pi_\gamma(\theta_{n-1})}\theta_{n-1} - \gamma \mathbb{E}_{\pi_\gamma(\theta_{n-1})}\mathbb{E}_{p(x_n,y_n)}f_n'(\theta_{n-1}), \end{aligned}$$

which leads to $\mathbb{E}_{\pi_\gamma(\theta)}\mathbb{E}_{p(x_n,y_n)}\nabla f_n(\theta) = 0$, that is, now removing the dependence on n (data (x, y) are i.i.d.):

$$\mathbb{E}_{\pi_\gamma(\theta)}\mathbb{E}_{p(x,y)}\Big[-y\Phi(x) + a'\big(\Phi(x)^\top\theta\big)\Phi(x)\Big] = 0,$$

which finally leads to

$$\mathbb{E}_{p(x)}\Big[\mathbb{E}_{\pi_\gamma(\theta)}a'\big(\Phi(x)^\top\theta\big) - \mu_{**}(x)\Big]\Phi(x) = 0. \quad (4)$$

This is the core equation our method relies on. It does not imply that $b(x) = \mathbb{E}_{\pi_\gamma(\theta)}a'\big(\Phi(x)^\top\theta\big) - \mu_{**}(x)$ is uniformly equal to zero (which we want), but only that $\mathbb{E}_{p(x)}\Phi(x)b(x) = 0$, i.e., $b(x)$ is uncorrelated with $\Phi(x)$.

If the feature vector $\Phi(x)$ is "large enough" then this is equivalent to $b = 0$.[1]

[1] Let $\Phi(x) = (\phi_1(x), \ldots, \phi_n(x))^\top$ be an orthogonal basis and $b(x) = \sum_{i=1}^n c_i\phi_i(x) + \varepsilon(x)$, where $\varepsilon(x)$ is small if the basis is big enough. Then $\mathbb{E}_{p(x)}\Phi(x)b(x) = 0 \Leftrightarrow \mathbb{E}\phi_i(x)\big[\sum_{i=1}^n c_i\phi_i(x) + \varepsilon(x)\big] = 0$ for every i, and due to the orthogonality of the basis and the smallness of $\varepsilon(x)$: $c_i \cdot \mathbb{E}_{p(x)}\phi^2(x) \approx 0$ and hence $c_i \approx 0$ and thus $b(x) \approx 0$.

For example, when $\Phi(x)$ is composed of an orthonormal basis of the space of integrable functions (like for kernels in Section 6), then this is exactly true. Thus, in this situation,

$$\mu_{**}(x) = \mathbb{E}_{\pi_\gamma(\theta)} a'\big(\Phi(x)^\top \theta\big), \tag{5}$$

and averaging predictions $a'\big(\Phi(x)^\top \theta_n\big)$, along the path (θ_n) of the Markov chain should exactly converge to the optimal prediction.

This exact convergence is weaker (requires high-dimensional fatures) than for the unconditional family in Section 3 but it can still bring surprising benefits even when Φ is not large enough, as we present in Section 5 and Section 6.

4.3 TWO TYPES OF AVERAGING

Now we can introduce two possible ways to estimate the prediction function $\mu(x)$.

Averaging estimators. The first one is the usual way: we first estimate parameter θ, using Ruppert-Polyak averaging (Polyak and Juditsky, 1992): $\bar{\theta}_n = \frac{1}{n+1}\sum_{i=0}^n \theta_i$ and then we denote

$$\bar{\mu}_n(x) = a'(\Phi(x)^\top \bar{\theta}_n) = a'\Big(\Phi(x)^\top \tfrac{1}{n+1}\textstyle\sum_{i=0}^n \theta_i\Big)$$

the corresponding prediction. As discussed in Section 2 it converges to $\bar{\mu}_\gamma : x \mapsto a'(\Phi(x)^\top \bar{\theta}_\gamma)$, which is *not* equal to in general to $a'(\Phi(x)^\top \theta_*)$, where θ_* is the optimal parameter in $\mathbb{R}^d$. Since, as presented at the end of Section 2, $\bar{\theta}_\gamma - \theta_*$ is of order $O(\gamma)$, $F(\bar{\theta}_\gamma) - F(\theta_*)$ is of order $O(\gamma^2)$ (because $\nabla F(\theta_*) = 0$), and thus an error of $O(\gamma^2)$ is added to the usual convergence rates in $O(1/n)$.

Note that we are limited here to prediction functions which corresponds to *linear functions* in $\Phi(x)$ in the natural parameterization, and thus $F(\theta_*) \geqslant \mathcal{G}(\mu_{**})$, and the inequality is often strict.

Averaging predictions. We propose a new estimator

$$\bar{\bar{\mu}}_n(x) = \frac{1}{n+1}\sum_{i=0}^n a'(\theta_i^\top \Phi(x)).$$

In general $\mathcal{G}(\bar{\bar{\mu}}_n) - \mathcal{G}(\mu_{**})$ does not converge to zero either (unless the feature vector Φ is large enough and Eq. (5) is satisfied). Thus, on top of the usual convergence in $O(1/n)$ with respect to the number of iterations, we have an extra term that depends only on γ, which we will study in Section 5 and Section 6.

We denote by $\bar{\bar{\mu}}_\gamma(x)$ the limit when $n \to \infty$, that is, using properties of converging Markov chains, $\bar{\bar{\mu}}_\gamma(x) = \mathbb{E}_{\pi_\gamma(\theta)} a'\big(\Phi(x)^\top \theta\big)$.

Rewriting Eq. (4) using our new notations, we get:

$$\mathbb{E}\big[(\mu_{**}(x) - \bar{\bar{\mu}}_\gamma(x))\Phi(x_n)\big] = 0.$$

When $\Phi : \mathbb{R} \to \mathbb{R}^d$ is high-dimensional, this leads to $\mu_{**} = \bar{\bar{\mu}}_\gamma$ and in contrast to $\bar{\mu}_\gamma$, averaging predictions potentially converge to the optimal prediction.

Computational complexity. Usual averaging of estimators (Polyak and Juditsky, 1992) to compute $\bar{\mu}_n(x) = a'(\Phi(x)^\top \bar{\theta}_n)$ is simple to implement as we can simply update the average $\bar{\theta}_n$ with essentially no extra cost on top the complexity $O(nd)$ of the SGD recursion. Given the number n of training data points and the number m of testing data points, the overall complexity is $O(d(n+m))$.

Averaging prediction functions is more challenging as we have to store all iterates θ_i, $i = 1, \dots, n$, and for each testing point x, compute $\bar{\bar{\mu}}_n(x) = \frac{1}{n+1}\sum_{i=0}^n a'(\theta_i^\top \Phi(x))$. Thus the overall complexity is $O(dn + mnd)$, which could be too costly with many test points (i.e., m large).

There are several ways to alleviate this extra cost: (a) using sketching techniques (Woodruff et al., 2014), (b) using summary statistics like done in applications of MCMC (Gilks et al., 1995), or (c) leveraging the fact that all iterates θ_i will end up being close to $\bar{\theta}_\gamma$ and use a Taylor expansion of $a'\big(\theta^\top \Phi(x)\big)$ around $\bar{\theta}_\gamma$. This expansion is equal to:

$$a'\big(\Phi(x)^\top \bar{\theta}_\gamma\big) + (\theta - \bar{\theta}_\gamma)^\top \Phi(x) \cdot a''\big(\Phi(x)^\top \bar{\theta}_\gamma\big) +$$

$$+ \frac{1}{2}\big((\theta - \bar{\theta}_\gamma)^\top \Phi(x)\big)^2 \cdot a'''\big(\Phi(x)^\top \bar{\theta}_\gamma\big) + O\big(\|\theta - \bar{\theta}_\gamma\|^3\big).$$

Taking expectation in both sides above leads to:

$$\bar{\bar{\mu}}_\gamma(x) \approx \bar{\mu}_\gamma(x) + \frac{1}{2}\Phi(x)^\top \mathrm{cov}\,(\theta) \cdot \Phi(x) \cdot a'''\big(\bar{\theta}_\gamma^\top \Phi(x)\big),$$

where $\mathrm{cov}\,(\theta)$ is the covariance matrix of θ under π_γ. This provides a simple correction to $\bar{\mu}_\gamma$, and leads to an approximation of $\bar{\bar{\mu}}_n(x)$ as

$$\bar{\mu}_n(x) + \frac{1}{2}\,\Phi(x)^\top \mathrm{cov}_n(\theta)\,\Phi(x) \cdot a'''\big(\bar{\theta}_n^\top \Phi(x)\big),$$

where $\mathrm{cov}_n(\theta)$ is the empirical covariance matrix of the iterates (θ_i).

The computational complexity now becomes $O(nd^2 + md^2)$, which is an improvement when the number of testing points m is large. In all of our experiments, we used this approximation.

5 FINITE-DIMENSIONAL MODELS

In this section we study the behavior of $\bar{\bar{A}}(\gamma) = \mathcal{G}(\bar{\bar{\mu}}_\gamma) - \mathcal{G}(\mu_*)$ for finite-dimensional models, for which it is usually not equal to zero. We know that our estimators $\bar{\bar{\mu}}_n$ will converge to $\bar{\bar{\mu}}_\gamma$, and our goal is to compare it to $\bar{A}(\gamma) = \mathcal{G}(\bar{\mu}_\gamma) - \mathcal{G}(\mu_*) = F(\bar{\theta}_\gamma) - F(\theta_*)$ which is what averaging estimators tends to. We also consider for completeness the non-averaged performance $A(\gamma) = \mathbb{E}_{\pi_\gamma(\theta)}\big[F(\theta) - F(\theta_*)\big]$.

Note that we must have $A(\gamma)$ and $\bar{A}(\gamma)$ non-negative, because we compare the negative log-likelihood performances to the one of of the best linear prediction (in the natural parameter), while $\bar{\bar{A}}(\gamma)$ could potentially be negative (it will in certain situations), because the the corresponding natural parameters may not be linear in $\Phi(x)$.

We consider the same standard assumptions as Dieuleveut et al. (2017), namely smoothness of the negative log-likelihoods $f_n(\theta)$ and strong convexity of the expected negative log-likelihood $F(\theta)$. We first recall the results from Dieuleveut et al. (2017). See detailed explicit formulas in the supplementary material.

5.1 EARLIER WORK

Without averaging. We have that $A(\gamma) = \gamma B + O(\gamma^{3/2})$, that is γ is *linear* in γ, with B non-negative.

Averaging estimators. We have that $\bar{A}(\gamma) = \gamma^2 \bar{B} + O(\gamma^{5/2})$, that is $\bar{A}$ is *quadratic* in γ, with $\bar{B}$ non-negative. Averaging does provably bring some benefits because the order in γ is higher (we assume γ small).

5.2 AVERAGING PREDICTIONS

We are now ready to analyze the behavior of our new framework of averaging predictions. The following result is shown in the supplementary material.

Proposition 1 *Under the assumptions on the negative loglikelihoods f_n of each observation from Dieuleveut et al. (2017):*

- *In the case of well-specified data, that is, there exists θ_* such that for all (x, y), $p(y|x) = q(y|x, \theta_*)$, then $\bar{\bar{A}} \sim \gamma^2 \bar{\bar{B}}^{\text{well}}$, where $\bar{\bar{B}}^{\text{well}}$ is a positive constant.*
- *In the general case of potentially mis-specified data, $\bar{\bar{A}} = \gamma \bar{\bar{B}}^{\text{mis}} + O(\gamma^2)$, where $\bar{\bar{B}}^{\text{mis}}$ is constant which may be positive or negative.*

Note, that in contrast to averaging parameters, the constant $\bar{\bar{B}}^{\text{mis}}$ can be negative. It means, that we obtain the estimator better than the optimal linear estimator, which is the limit of capacity for averaging parameters. In our simulations, we show examples for which $\bar{\bar{B}}^{\text{mis}}$ is positive, and examples for which it is negative. Thus, in general, for low-dimensional models, averaging predictions can be worse or better than averaging parameters. However, as we show in the next section, for infinite dimensional models, we always get convergence.

6 INFINITE-DIMENSIONAL MODELS

Recall, that we have the following objective function to minimize:

$$F(\theta) = \mathbb{E}_{x,y}\Big[- y \cdot \eta_\theta(x) + a\big(\eta_\theta(x)\big)\Big], \qquad (6)$$

where till this moment we consider unknown functions $\eta_\theta(x)$ which were linear in $\Phi(x)$ with $\Phi(x) \in \mathbb{R}^d$, leading to a complexity in $O(dn)$.

We now consider infinite-dimensional features, by considering that $\Phi(x) \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space. Note that this corresponds to modeling the function η_θ as a Gaussian process (Rasmussen and Williams, 2006).

This is computationally feasible through the usual "kernel trick" (Scholkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004), where we assume that the kernel function $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ is easy to compute. Indeed, following Bordes et al. (2005) and Dieuleveut and Bach (2016), starting from θ_0, each iterate of constant-step-size SGD is of the form $\theta_n = \sum_{t=1}^n \alpha_t \Phi(x_t)$, and the gradient descent recursion $\theta_n = \theta_{n-1} - \gamma[a'(\eta_{\theta_{n-1}}(x_n)) - y_n]\Phi(x_n)$ leads to the following recursion on α_t's:

$$\begin{aligned} \alpha_n &= -\gamma\big[a'\big(\textstyle\sum_{t=1}^{n-1} \alpha_t \langle \Phi(x_t), \Phi(x_n) \rangle\big) - y_n\big] \\ &= -\gamma\big[a'\big(\textstyle\sum_{t=1}^{n-1} \alpha_t k(x_t, x_n)\big) - y_n\big]. \end{aligned}$$

This leads to $\eta_{\theta_n}(x) = \langle \Phi(x), \theta_n \rangle$ and $\mu_{\theta_n}(x) = a'\big(\eta_{\theta_n}(x)\big)$ with

$$\eta_{\theta_n}(x) = \sum_{t=1}^n \alpha_t \langle \Phi(x), \Phi(x_t) \rangle = \sum_{t=1}^n \alpha_t k(x, x_t),$$

and finally we can express $\bar{\bar{\mu}}_n(x)$ in kernel form as:

$$\bar{\bar{\mu}}_n(x) = \frac{1}{n+1} \sum_{t=0}^n a'\Big[\sum_{l=1}^t \alpha_l \cdot k(x, x_l)\Big].$$

There is also a straightforward estimator for averaging parameters, i.e., $\bar{\mu}_n(x) = a'\Big(\frac{1}{n+1} \sum_{t=0}^n \sum_{l=1}^t \alpha_l k(x, x_l)\Big)$. If we assume that the kernel function is *universal*, that is, $\mathcal{H}$

is dense in the space of squared integrable functions, then it is known that if $\mathbb{E}_x b(x)\Phi(x) = 0$, then $b = 0$ (Sriperumbudur et al., 2008). This implies that we must have $\bar{\mu}_\gamma = 0$ and thus averaging predictions does always converge to the global optimum (note that in this setting, we must have a well-specified model because we are in a non-parametric setting).

Column sampling. Because of the usual high running-time complexity of kernel method in $O(n^2)$, we consider a "column-sampling approximation" (Williams and Seeger, 2001). We thus choose a small subset $I = (x_1, \dots, x_m)$ of samples and construct a new finite m-dimensional feature map $\bar{\Phi}(x) = K(I,I)^{-1/2}K(I,x) \in \mathbb{R}^m$, where $K(I,I)$ is the $m \times m$ kernel matrix of the m points and $K(I,x)$ the vector composed of kernel evaluations $k(x_i, x)$. This allows a running-time complexity in $O(m^2 n)$. In theory and practice, m can be chosen small (Bach, 2013; Rudi et al., 2017).

Regularized learning with kernels. Although we can use an unregularized recursion with good convergence properties (Dieuleveut and Bach, 2016), adding a regularisation by the squared Hilbertian norm is easier to analyze and more stable with limited amounts of data. We thus consider the recursion (in Hilbert space), with λ small:

$$\begin{aligned}\theta_n &= \theta_{n-1} - \gamma\big[f_n'(\theta_{n-1}) + \lambda\theta_{n-1}\big] \\ &= \theta_{n-1} + \gamma(y_n - a'(\langle\Phi(x_n), \theta\rangle))\Phi(x_n) - \gamma\lambda\theta_{n-1}.\end{aligned}$$

This recursion can also be computed efficiently as above using the kernel trick and column sampling approximations.

In terms of convergence, the best we can hope for is to converge to the minimizer $\theta_{*,\lambda}$ of the regularized expected negative log-likelihood $F(\theta) + \frac{\lambda}{2}\|\theta\|^2$ (which we assume to exist). When λ tends to zero, then $\theta_{*,\lambda}$ converges to θ_*.

Averaging *parameters* will tend to a limit $\bar{\theta}_{\gamma,\lambda}$ which is $O(\gamma)$-close to $\theta_{*,\lambda}$, thus leading to predictions which deviate from the optimal predictions for two reasons: because of regularization and because of the constant stepsize. Since λ should decrease as we get more data, the first effect will vanish, while the second will not.

When averaging *predictions*, the two effects will vanish as λ tends to zero. Indeed, by taking limits of the gradient equation, and denoting by $\bar{\bar{\mu}}_{\gamma,\lambda}$ the limit of $\bar{\bar{\mu}}_n$, we have

$$\mathbb{E}\big[(\mu_{**}(x) - \bar{\bar{\mu}}_{\gamma,\lambda}(x))\Phi(x)\big] = \lambda\bar{\theta}_{\gamma,\lambda}. \qquad (7)$$

Given that $\bar{\theta}_{\gamma,\lambda}$ is $O(\gamma)$-away from θ_*, if we assume that θ_* corresponds to a sufficiently regular[2] element of the Hilbert space $\mathcal{H}$, then the L_2-norm of the deviation satisfies $\|\mu_{**}(x) - \bar{\bar{\mu}}_{\gamma,\lambda}\| = O(\lambda)$ and thus as the regularization parameter λ tends to zero, our predictions tend to the optimal one.

7 EXPERIMENTS

In this section, we compare the two types of averaging (estimators and predictions) on a variety of problems, both on synthetic data and on standard benchmarks. When averaging predictions, we always consider the Taylor expansion approximation presented at the end of Section 4.3.

7.1 SYNTHETIC DATA

Finite-dimensional models. we consider the following logistic regression model:

$$q(y|x,\theta) = \exp\big(y \cdot \eta_\theta(x) - a(\eta_\theta(x))\big),$$

where we consider a linear model $\eta_\theta(x) = \theta^\top x$ in x (i.e., $\Phi(x) = x$), the link function $a(t) = \log(1+e^t)$ and $a'(t) = \sigma(t)$ is the sigmoid function. Let x be distributed as a standard normal random variable in dimension $d = 2$, $y \in \{0,1\}$ and $\mathbb{P}(y = 1|x) = \mu_{**}(x) = \sigma\big(\eta_{**}(x)\big)$, where we consider two different settings:

- Model 1: $\eta_{**}(x) = \sin x_1 + \sin x_2$,
- Model 2: $\eta_{**}(x) = x_1^3 + x_2^3$.

The global minimum $\mathcal{F}_{**}$ of the corresponding optimization problem can be found as

$$\mathcal{F}_{**} = \mathbb{E}_{p(x)}\big[-\mu_{**}(x) \cdot \eta_{**}(x) + a(\eta_{**}(x))\big].$$

We also introduce the performance measure $\mathcal{F}(\eta)$

$$\mathcal{F}(\eta) = \mathbb{E}_{p(x)}\big[-\mu_{**}(x) \cdot \eta(x) + a(\eta(x))\big], \qquad (8)$$

which can be evaluated directly in the case of synthetic data. Note that in our situation, the model is misspecified because $\eta_{**}(x)$ is not linear in $\Phi(x) = x$, and thus, $\inf_\theta F(\theta) > \mathcal{F}_{**}$, and thus our performance measures $\mathcal{F}(\mu_n) - \mathcal{F}_{**}$ for various estimators μ_n will not converge to zero.

The results of averaging 10 replications are shown in Fig. 3 and Fig. 4. We first observed that constant stepsize SGD without averaging leads to a bad performance.

[2] While our reasoning is informal here, it can be made more precise by considering so-called "source conditions" commonly used in the analysis of kernel methods (Caponnetto and De Vito, 2007), but this is out of the scope of this paper.

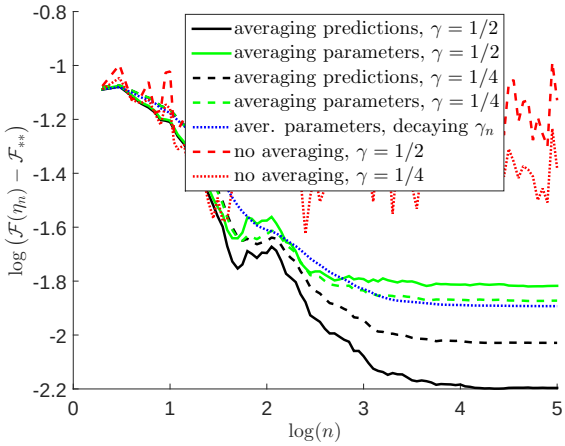

Figure 3: Synthetic data for linear model $\eta_\theta(x) = \theta^\top x$ and $\eta_{**}(x) = \sin x_1 + \sin x_2$. Excess prediction performance vs. number of iterations (both in log-scale).

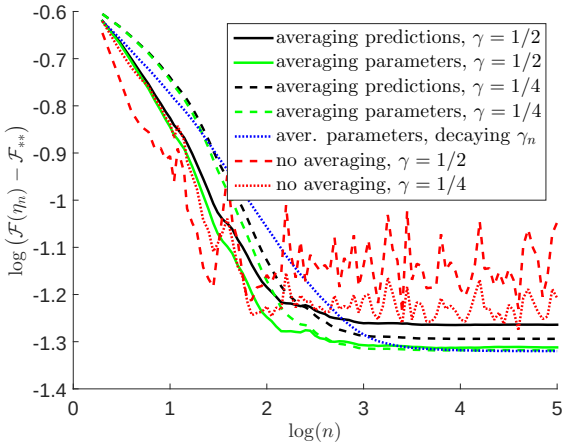

Figure 4: Synthetic data for linear model $\eta_\theta(x) = \theta^\top x$ and $\eta_{**}(x) = x_1^3 + x_2^3$. Excess prediction performance vs. number of iterations (both in log-scale).

Moreover, we can see, that in the first case (Fig. 3) averaging predictions beats averaging parameters, and moreover beats the best linear model: if we use the best linear error $\mathcal{F}_*$ instead of $\mathcal{F}_{**}$, at some moment $\mathcal{F}(\eta_n) - \mathcal{F}_*$ becomes negative. However in the second case (Fig. 4), averaging predictions is not superior to averaging parameters. Moreover, by looking at the final differences between performances with different values of γ, we can see the dependency of the final performance in γ for averaging predictions, instead of γ^2 for averaging parameters (as suggested by our theoretical results in Section 5). In particular in Fig. 3, we can observe the surprising behavior of a larger step-size leading to a better performance (note that we cannot increase too much otherwise the algorithm would diverge).

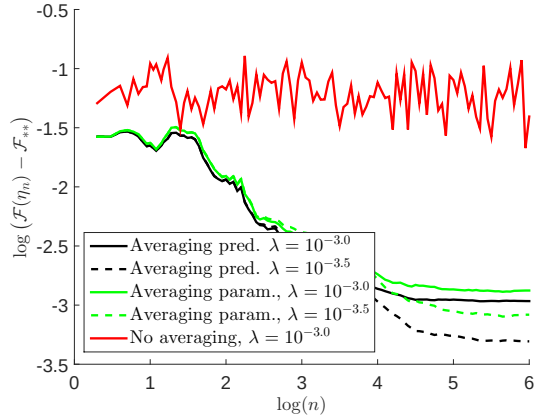

Figure 5: Synthetic data for Laplacian kernel for $\eta_{**}(x) = \frac{5}{5+x^\top x}$. Excess prediction performance vs. number of iterations (both in log-scale).

Infinite-dimensional models Here we consider the kernel setup described in Section 6. We consider Laplacian kernels $k(s,t) = \exp\left(\frac{\|s-t\|_1}{\sigma}\right)$ with $\sigma = 50$, dimension $d = 5$ and generating log odds ratio $\eta_{**}(x) = \frac{5}{5+x^\top x}$. We also use a squared norm regularization with several values of λ and column sampling with $m = 100$ points. We use the exact value of $\mathcal{F}_{**}$ which we can compute directly for synthetic data. The results are shown in Fig. 5, where averaging predictions leads to a better performance than averaging estimators.

7.2 REAL DATA

Note, that in the case of real data, one does not have access to unknown $\mu_{**}(x)$ and computing the performance measure in Eq. (8) is inapplicable. Instead of it we use its sampled version on held out data:

$$\hat{\mathcal{F}}(\eta) = -\sum_{i:y_i=1} \log\left(\mu(x_i)\right) - \sum_{j:y_j=0} \log\left(1-\mu(x_i)\right).$$

We use two datasets, with d not too large, and n large, from (Lichman, 2013): the "MiniBooNE particle identification" dataset ($d = 50$, $n = 130\,064$), the "Covertype" dataset ($d = 54$, $n = 581\,012$).

We use two different approaches for each of them: a linear model $\eta_\theta(x) = \theta^\top x$ and a kernel approach with Laplacian kernel $k(s,t) = \exp\left(\frac{\|s-t\|_1}{\sigma}\right)$, where $\sigma = d$. The results are shown in Figures 6 to 9. Note, that for linear models we use $\hat{\mathcal{F}}_*$–the estimator of the best performance among linear models (learned on the test set, and hence not reachable from learning on the training data), and for kernels we use $\hat{\mathcal{F}}_{**}$ (same definition as $\hat{\mathcal{F}}_*$ but with the kernelized model), that is why graphs are not comparable (but, as shown below, the value of $\hat{\mathcal{F}}_{**}$ is

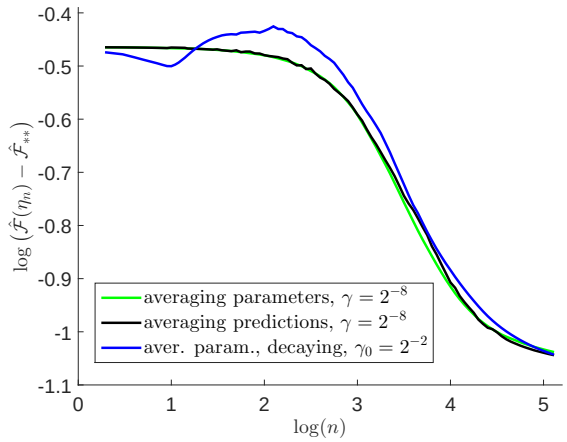

Figure 6: MiniBooNE dataset, dimension $d = 50$, linear model. Excess prediction performance vs. number of iterations (both in log-scale).

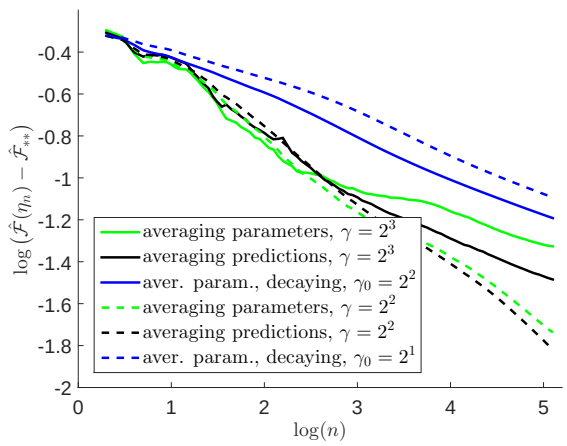

Figure 7: MiniBooNE dataset, dimension $d = 50$, kernel approach, column sampling $m = 200$. Excess prediction performance vs. number of iterations (both in log-scale).

lower than the value of $\hat{\mathcal{F}}_*$ because using kernels correspond to a larger feature space).

For the "MiniBooNE particle identification" dataset $\hat{\mathcal{F}}_* = 0.35$ and $\hat{\mathcal{F}}_{**} = 0.21$; for the"Covertype" dataset $\hat{\mathcal{F}}_* = 0.46$ and $\hat{\mathcal{F}}_{**} = 0.39$. We can see from the four plots that, especially in the kernel setting, averaging predictions also shows better performance than averaging parameters.

8 CONCLUSION

In this paper, we have explored how averaging procedures in stochastic gradient descent, which are crucial for fast convergence, could be improved by looking at the specifics of probabilistic modeling. Namely, averaging

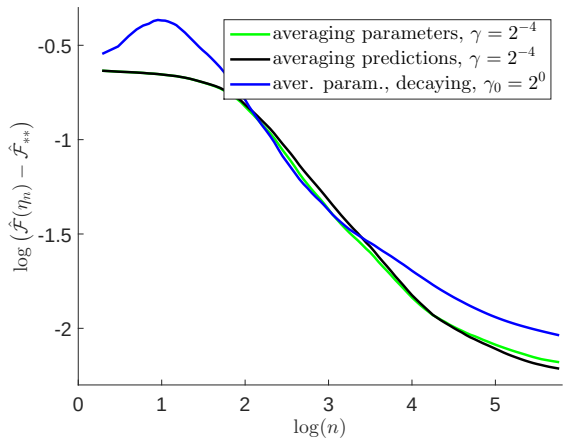

Figure 8: CoverType dataset, dimension $d = 54$, linear model. Excess prediction performance vs. number of iterations (both in log-scale).

in the moment parameterization can have better properties than averaging in the natural parameterization.

While we have provided some theoretical arguments (asymptotic expansion in the finite-dimensional case, convergence to optimal predictions in the infinite-dimensional case), a detailed theoretical analysis with explicit convergence rates would provide a better understanding of the benefits of averaging predictions.

Acknowledgements

The research leading to these results has received funding from the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement n° 642685 MacSeNet, and from the European Research Council (grant SEQUOIA 724063).

References

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Adv. NIPS*, 2011.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6(Sep):1579–1619, 2005.

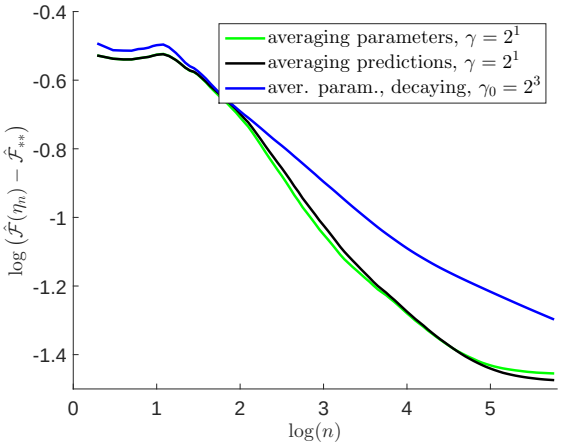

Figure 9: CoverType dataset, dimension $d = 54$, kernel approach, column sampling $m = 200$. Excess prediction performance vs. number of iterations (both in log-scale).

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. Technical Report 1606.04838, arXiv, 2016.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4): 1363–1399, 08 2016.

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. Technical Report 1707.06386, arXiv, 2017.

W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

J. M. Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, 2001.

M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

P. McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer-Verlag Inc, Berlin; New York, 1993.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

A. Rudi, L. Carratino, and L. Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3891–3901, 2017.

B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*. MIT press, 2001.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge university press, 2004.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective hilbert space embeddings of probability measures. In *Proc. COLT*, 2008.

C. K.I. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.

D. P. Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.