Sequential Analysis

Design Methods and Applications

ISSN: 0747-4946 (Print) 1532-4176 (Online) Journal homepage: https://www.tandfonline.com/loi/lsqa20

Adaptive sequential machine learning

Craig Wilson, Yuheng Bu & Venugopal V. Veeravalli

To cite this article: Craig Wilson, Yuheng Bu & Venugopal V. Veeravalli (2019) Adaptive sequential machine learning, Sequential Analysis, 38:4, 545-568, DOI: 10.1080/07474946.2019.1686889

To link to this article: https://doi.org/10.1080/07474946.2019.1686889

Published online: 29 Jan 2020.

SEQUENTIAL ANALYSIS
2019, VOL. 38, NO. 4, 545–568
https://doi.org/10.1080/07474946.2019.1686889

Adaptive sequential machine learning

Craig Wilson, Yuheng Bu, and Venugopal V. Veeravalli

ECE Department and Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, Illinois, USA

ABSTRACT

A framework previously introduced in Wilson et al. (2018) for solving a sequence of stochastic optimization problems with bounded changes in the minimizers is extended and applied to machine learning problems such as regression and classification. The stochastic optimization problems arising in these machine learning problems are solved using algorithms such as stochastic gradient descent (SGD). A method based on estimates of the change in the minimizers and properties of the optimization algorithm is introduced for adaptively selecting the number of samples at each time step to ensure that the excess risk—that is, the expected gap between the loss achieved by the approximate minimizer produced by the optimization algorithm and the exact minimizer—does not exceed a target level. A bound is developed to show that the estimate of the change in the minimizers is non trivial provided that the excess risk is small enough. Extensions relevant to the machine learning setting are considered, including a cost-based approach to select the number of samples with a cost budget over a fixed horizon, and an approach to applying cross-validation for model selection. Finally, experiments with synthetic and real data are used to validate the algorithms.

ARTICLE HISTORY
Received 21 August 2019
Revised 10 September 2019
Accepted 15 September 2019

KEYWORDS
Adaptive learning; cross-validation; excess risk; sequential learning; stochastic gradient descent

SUBJECT CLASSIFICATIONS
68T05; 62L05; 62L10

1. Introduction

Consider solving a sequence of machine learning problems by minimizing the following expected value of a fixed loss function $\ell(\boldsymbol{w}, \boldsymbol{z})$ at each time n:

$$\min_{\boldsymbol{w}\in\mathcal{X}}\{f_n(\boldsymbol{w}) \triangleq \mathbb{E}_{\boldsymbol{z}_n\sim p_n}[\ell(\boldsymbol{w}, \boldsymbol{z}_n)]\}, \quad \forall n \geq 1, \tag{1.1}$$

where p_n denotes the underlying (unknown) probabilistic model for the data $\boldsymbol{z}_n$ at time n and $\mathcal{X}$ is a closed and convex parameter space. For regression, $\boldsymbol{z}_n = \{\boldsymbol{x}_n, y_n\}$ corresponds to the {predictors, response} pair at time n and $\boldsymbol{w}$ parameterizes the regression model. For classification, $\boldsymbol{z}_n = \{\boldsymbol{x}_n, y_n\}$ corresponds to the {features, label} pair at time n, and $\boldsymbol{w}$ parameterizes the classifier. Although motivated by regression and classification, our framework works for any loss function $\ell(\boldsymbol{w}, \boldsymbol{z})$ that satisfies certain properties discussed in Section 2.1.

CONTACT Venugopal V. Veeravalli ✉ vvv@illinois.edu ECE Department, University of Illinois at Urbana–Champaign, 315 Coordinated Science Laboratory, 1308 West Main Street, Urbana, IL 61801, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsqa.
Recommended by Edit Gombay

We assume that problem (1.1) has a unique solution, denoted by $\boldsymbol{w}_n^*$ at every instance n; that is,

$$\boldsymbol{w}_n^* \triangleq \arg\min_{\boldsymbol{w}\in\mathcal{X}} f_n(\boldsymbol{w}), \quad \forall n \geq 1. \tag{1.2}$$

By imposing a condition on the minimizers $\boldsymbol{w}_n^*$ of the function $f_n(\boldsymbol{w})$, we assume that these machine learning problems change at a bounded but unknown rate:

$$||\boldsymbol{w}_n^* - \boldsymbol{w}_{n-1}^*|| \leq \rho, \quad \forall n \geq 2, \tag{1.3}$$

where we use $||\cdot||$ to denote ℓ_2 norm and ρ is a finite upper bound on the change of minimizers, which needs to be estimated in practice.

Under this model, we find approximate minimizers $\boldsymbol{w}_n$ of each function $f_n(\boldsymbol{w})$ by drawing K_n samples $\{\boldsymbol{z}_n(k)\}_{k=1}^{K_n} \overset{\text{iid}}{\sim} p_n$ at time n. We do not make any assumptions about the particular optimization algorithm that may be used to find the approximate minimizers. As an example, we could use these samples in an optimization algorithm such as stochastic gradient descent (SGD). We evaluate the quality of our approximate minimizers $\boldsymbol{w}_n$ through an excess risk criterion with level ϵ; that is,

$$\mathbb{E}[f_n(\boldsymbol{w}_n)] - f_n(\boldsymbol{w}_n^*) \leq \epsilon, \tag{1.4}$$

which is a standard criterion for optimization and learning problems (see, e.g., Mohri et al., 2012). Note that the expectation is taken over the randomness of the approximate minimizers $\boldsymbol{w}_n$. Our goal is to determine *adaptively* the number of samples K_n required to achieve a desired excess risk ϵ for large enough n. Because ρ is unknown, we will first construct an estimate of ρ. Given an estimate of ρ, we determine selection rules for the number of samples K_n to achieve a target excess risk ϵ.

This article is a continuation of the work initiated in Wilson et al. (2018). We specialize the results in Wilson et al. (2018), which were given for general functions $f_n(\boldsymbol{w})$, to the specific form in (1.1) and provide new results that are specifically relevant to machine learning problems. We develop a bound to show that our estimate ρ is nontrivial provided that the excess risk is small enough. We also consider extensions relevant to the machine learning setting, including a cost-based approach to select the number of samples with a cost budget over a fixed horizon and an approach to applying cross-validation for model selection. Some of the results in this paper have reported in conference publications (Wilson and Veeravalli, 2016a,b), which do not contain proofs of the key results due to space limitations. Moreover, we provide substantially more detailed numerical results and simulations in this article than those given in Wilson and Veeravalli (2016a,b).

1.1. Related work

Our problem has connections with *multitask learning* (MTL) and *transfer learning*. In MTL, one tries to learn several tasks simultaneously as in Agarwal et al. (2011), Evgeniou and Pontil (2004), and Zhang and Yeung (2012) by exploiting the relationships between the tasks. In transfer learning, knowledge from one source task is transferred to another target task either with or without additional training data for the target task as in Pan and Yang (2010). For multitask and transfer learning, there are theoretical guarantees on regret for some algorithms (e.g., Agarwal et al., 2008). MTL could be applied to our problem by running an MTL algorithm each time a new task arrives, while remembering all prior tasks. However, this approach incurs a memory and computational burden. Transfer learning lacks the sequential nature of our problem.

We can also consider the *concept drift* problem in which we observe a stream of incoming data that potentially changes over time, and the goal is to predict some property of each piece of data as it arrives. After prediction, we incur a loss that is revealed to us. For example, we could observe a feature $\boldsymbol{x}_n$ and predict the label y_n as in Towfic et al. (2013). Some approaches for concept drift use iterative algorithms such as SGD but without specific models on how the data changes. As a result, only simulation results showing good performance are available.

Another related problem is online optimization, where generally no knowledge is available about the incoming functions other than that all of the functions come from a specified class of functions; that is, linear or convex functions with uniformly bounded gradients. Online optimization models do not include the notion of a desired excess risk bound. Rather, only bounds on the regret over some time horizon have been investigated, as in Cesa-Bianchi and Lugosi (2006), Duchi et al. (2011), Duchi and Singer (2009), Hazan et al. (2007), Bartlett et al. (2007), Shalev-Shwartz and Kakade (2009), Shalev-Shwartz and Singer (2006, 2007), Xiao (2010), and Zinkevich (2003), which is different from the per time step excess risk guarantee provided in our work.

There has been some work on controlling the variation of the sequence of functions $f_n(\boldsymbol{w})$ in (1.1) in Rakhlin and Sridharan (2012) and Chiang et al. (2012). The work in Chiang et al. (2012) is most relevant where regret is minimized subject to a bound, say G_b, on the total variation of the gradients over a time interval T of interest; that is,

$$\sum_{n=1}^{T} \max_{\boldsymbol{w}\in\mathcal{X}} ||\nabla f_{n+1}(\boldsymbol{w}) - \nabla f_n(\boldsymbol{w})||^2 \leq G_b. \tag{1.5}$$

If all of the functions $\{f_n(x)\}$ are strongly convex with the same parameter m (See Assumption A.2 in Section 2.1 for the definition of strong convexity), then by the optimality conditions (see theorem 2F.10 in Dontchev and Rockafellar, 2009) (1.5) implies that

$$\sum_{n=1}^{T} ||\boldsymbol{w}^*_{n+1} - \boldsymbol{w}^*_n||^2 \leq \frac{G_b}{m^2}.$$

Thus, the work in Chiang et al. (2012) can be seen as studying the regret with a constraint on the total variation in the minimizers over T time instants. In contrast, we control the variation of the minimizers at each time instant with (1.3) and then seek to maintain an excess risk criterion such as (1.4) at each time step.

Another relevant model is *sequential supervised learning* (see Dietterich, 2002) in which we observe a stream of data consisting of feature/label pairs $(\boldsymbol{x}_n, y_n)$ at time n, with $\boldsymbol{x}_n$ being the feature vector and y_n being the label. At time n, we want to predict y_n given $\boldsymbol{x}_n$. One approach to this problem, studied in Fawcett and Provost (1997) and Qian and Sejnowski (1988), is to look at L consecutive pairs $\{(\boldsymbol{x}_{n-i}, y_{n-i})\}_{i=1}^{L}$ and develop a predictor at time n by applying a supervised learning algorithm to these training data. Another approach is to assume that there is an underlying hidden Markov model governing the data as in Bengio and Frasconi (1996). The label y_n represents the hidden state and the pair $(\boldsymbol{x}_n, \bar{y}_n)$ represents the observation with $\bar{y}_n$ being a noisy version of y_n. Hidden Markov model inference techniques are used to estimate y_n.

To summarize, none of the prior work discussed in this section involves adaptively choosing the number of samples K_n at each time n to control the excess risk. Most approaches instead focus on bounding the regret or provide no guarantees.

1.2. Article outline

The rest of this article is outlined as follows. In Section 2, we specialize the work in Wilson et al. (2018) to the machine learning problems. We introduce a method from Wilson et al. (2018) to estimate the unknown change ρ and establish the excess risk guarantees for the sequence of learning problems in (1.1). In Section 3, we develop an upper bound on the size of the overshoot of our estimate of ρ above the true value of ρ. In Section 4, we consider a cost-based approach to select the number of samples based on the analysis in Section 2, and a cross-validation approach. Finally, in Section 5, we apply our framework to a variety of machine learning problems on both synthetic and real data.

2. Adaptive sequential optimization

We summarize our previous work in Wilson et al. (2018) and apply it to the machine learning problem stated in (1.1).

2.1. Assumptions

We make several assumptions to proceed. First, let $\mathcal{X}$ be closed and convex with $\operatorname{diam}(\mathcal{X}) < +\infty$. Define the σ-algebra

$$\mathcal{F}_i \triangleq \sigma\Big(\{\boldsymbol{z}_j(k) : j = 1, \ldots, i; k = 1, \ldots, K_j\}\Big), \tag{2.1}$$

which is the smallest σ-algebra such that the random variables in the set $\{\boldsymbol{z}_j(k) : j = 1, \ldots, i; k = 1, \ldots, K_j\}$ are measurable. By convention $\mathcal{F}_0$ is the trivial σ-algebra.

We suppose that the following conditions hold:

A.1 For each n, $f_n(\boldsymbol{w})$ is twice continuously differentiable with respect to $\boldsymbol{w}$.
A.2 For each n, $f_n(\boldsymbol{w})$ is strongly convex with a parameter $m > 0$; that is,

$$f_n(\tilde{\boldsymbol{w}}) \geq f_n(\boldsymbol{w}) + \langle \nabla f_n(\boldsymbol{w}), \tilde{\boldsymbol{w}} - \boldsymbol{w} \rangle + \frac{1}{2} m \|\tilde{\boldsymbol{w}} - \boldsymbol{w}\|^2, \tag{2.2}$$

where $\langle \boldsymbol{w}, \tilde{\boldsymbol{w}} \rangle$ is the Euclidean inner product between $\boldsymbol{w}, \tilde{\boldsymbol{w}} \in \mathcal{X}$.
A.3 For each n, we can draw stochastic gradients $\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{z}_n(k))$, where $\{\boldsymbol{z}_n(k)\}_{k=1}^{K_n} \overset{\text{iid}}{\sim} p_n$ and

$$\mathbb{E}\Big[\nabla_{\boldsymbol{w}} \ell\Big(\boldsymbol{w}, \boldsymbol{z}_n(k)\Big)\Big] = \nabla f_n(\boldsymbol{w}). \tag{2.3}$$

A.4 Given an optimization algorithm that generates an approximate minimizer

$$\boldsymbol{w}_n \triangleq \mathscr{A}(\boldsymbol{w}_{n-1}, \{\boldsymbol{z}_n(k)\}_{k=1}^{K_n}) \tag{2.4}$$

using K_n samples $\{\boldsymbol{z}_n(k)\}_{k=1}^{K_n}$, there exists a function $b(d_0, K_n)$ such that the following conditions hold:

1. If K_n and d_0 are both $\mathscr{F}_{n-1}$-measurable random variables, it holds that
$$||\boldsymbol{w}_{n-1} - \boldsymbol{w}_n^*||^2 \le d_0^2 \Rightarrow \mathbb{E}\left[f_n(\boldsymbol{w}_n)|\mathscr{F}_{n-1}\right] - f_n(\boldsymbol{w}_n^*) \le b(d_0, K_n). \tag{2.5}$$
2. The bound $b(d_0, K_n)$ is non-decreasing in d_0 and non-increasing in K_n.

A.5 Initial approximate minimizers $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ satisfy

$$f_i(\boldsymbol{w}_i) - f_i(\boldsymbol{w}_i^*) \le \epsilon_i, i = 1, 2$$

with ϵ_1 and ϵ_2 known.

We note that Assumption A.3 guarantees that the gradients used in the optimization algorithm are unbiased. In addition, we assume that the bound $b(d_0, K_n)$ depends on the number of samples K_n and not the number of iterations in Assumption A.4. For the basic version of SGD, generally the number of iterations equals K_n, because each sample is used to produce a noisy gradient. See Wilson et al. (2018) for a thorough discussion of useful $b(d_0, K_n)$ bounds. For some $b(d_0, K)$, we may need to know parameters such as the strong convexity parameter. Estimating these parameters is also discussed in Wilson et al. (2018). Finally, for Assumption A.5, we can fix K_i and set $\epsilon_i = b(\text{diam}(\mathscr{X}), K_i)$ for $i = 1, 2$ at initialization.

2.2. Change in minimizers known

Following Wilson et al. (2018), we examine the case when the change in minimizers ρ is known. Suppose that ϵ_{n-1} bounds the excess risk at time n – 1. Using the triangle inequality, strong convexity, Jensen's inequality, and (1.3), we have

$$\mathbb{E}||\boldsymbol{w}_{n-1} - \boldsymbol{w}_n^*||^2 \le \left(\sqrt{\frac{2\epsilon_{n-1}}{m}} + \rho\right)^2. \tag{2.6}$$

Now, by using the bound $b(d_0, K_n)$ from Assumption A.4, we set

$$\epsilon_n = b\left(\sqrt{\frac{2\epsilon_{n-1}}{m}} + \rho, K_n\right), \forall n \ge 3, \tag{2.7}$$

yielding a sequence of bounds on the excess risk. Note that this recursion only relies on the immediate past at time n – 1 through ϵ_{n-1}. To achieve $\epsilon_n \le \epsilon$ for all n, we set

$$K_1 = \min\left\{K \ge 1 \,\middle|\, b\Big(\text{diam}(\mathscr{X}), K\Big) \le \epsilon\right\}$$

and $K_n = K^*$ for $n \ge 2$, with

$$K^* = \min\left\{K \ge 1 \,\middle|\, b\left(\sqrt{\frac{2\epsilon}{m}} + \rho, K\right) \le \epsilon\right\}. \tag{2.8}$$

In comparison, if we did not exploit the fact that the change is bounded by ρ, we would use the estimate $\operatorname{diam}^2(\mathcal{X})$ to bound $\mathbb{E}||\boldsymbol{w}_{n-1} - \boldsymbol{w}_n^*||^2$ and select K_n. If the bound in (2.6) is smaller than $\operatorname{diam}^2(\mathcal{X})$, then we would need significantly fewer samples K_n to guarantee a desired excess risk.

2.3. K^* may be too large

In this section, we look at a case where K^* can be too large. Suppose that $\rho = 0$, so the problems are not changing. In this case, we only need to take training samples at the first time instant and then we can stop taking samples; that is, $K_1 > 0$ and $K_n = 0$ for $n > 1$.

Suppose that $\epsilon_1 \leq \epsilon$ and $\rho = 0$. In this case, from the analysis in the previous section, we pick

$$K^* = \min\left\{K \geq 1 \;\middle|\; b\left(\sqrt{\frac{2\epsilon}{m}} + \rho, K\right) \leq \epsilon\right\}.$$

This implies that $K^* > 0$. However, by picking $K_n = 0$ for all $n \geq 2$, we could achieve $\epsilon_n = \epsilon_1 \leq \epsilon$ for all $n \geq 2$. This shows that the choice of K^* is conservative and can be too large if the initial distance $d_0 = 0$.

For an algorithm like SGD, the bound $b(d_0, K)$ is roughly of the form (see Wilson et al., 2018):

$$b(d_0, K) \approx \frac{1}{K} + \frac{d_0^2}{K^2}.$$

The first term captures the asymptotic behavior of SGD and the second term accounts for the initial distance d_0. As a general rule, the choice of K^* is useful if the term that depends on the initial distance, d_0^2/K^2, is comparable to the asymptotic term, $1/K$, in the $b(d_0, K)$ bound.

2.4. Estimating the change in the minimizers

In practice, we do not know ρ, so we must construct an estimate $\hat{\rho}_n$ using the samples $\{z_n(k)\}_{k=1}^{K_n}$ from each distribution p_n. We introduce an approach to estimate the one–time step change, $||\boldsymbol{w}_i^* - \boldsymbol{w}_{i-1}^*||$ and a method to combine these estimates to produce an overall estimate of ρ. These estimates are from Wilson et al. (2018). For appropriately chosen sequences $\{t_n\}$ and for all n large enough, we have $\hat{\rho}_n + t_n \geq \rho$ almost surely. With this property, analysis similar to that in Section 2.2 holds, which is provided in Section 2.5.

2.4.1. Estimating one-step change

First, we develop an estimate $\tilde{\rho}_i$ of the one-step changes $||\boldsymbol{w}_i^* - \boldsymbol{w}_{i-1}^*||$ using a method from Wilson et al. (2018). Implicitly, we assume that all one-step estimates are bounded by $\operatorname{diam}(\mathcal{X})$, because trivially $||\boldsymbol{w}_n^* - \boldsymbol{w}_{n-1}^*|| \leq \operatorname{diam}(\mathcal{X})$.

Using the triangle inequality and variational inequalities from Dontchev and Rockafellar (2009) yields

$$\begin{aligned}||\boldsymbol{w}_i^* - \boldsymbol{w}_{i-1}^*|| &\le ||\boldsymbol{w}_i - \boldsymbol{w}_{i-1}|| + ||\boldsymbol{w}_i - \boldsymbol{w}_i^*|| + ||\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-1}^*|| \\ &\le ||\boldsymbol{w}_i - \boldsymbol{w}_{i-1}|| + \frac{1}{m}||\nabla f_i(\boldsymbol{w}_i)|| + \frac{1}{m}||\nabla f_i(\boldsymbol{w}_{i-1})||.\end{aligned}$$

We then approximate $||\nabla f_i(\boldsymbol{w}_i)|| = ||\mathbb{E}_{\boldsymbol{z}_i \sim p_i}[\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w}_i, \boldsymbol{z}_i)]||$ by a sample average approximation to yield the following estimate called the *direct estimate*:

$$\tilde{\rho}_i \triangleq ||\boldsymbol{w}_i - \boldsymbol{w}_{i-1}|| + \frac{1}{m}||\hat{G}_i|| + \frac{1}{m}||\hat{G}_{i-1}||, \tag{2.9}$$

where

$$\hat{G}_i \triangleq \frac{1}{K_i}\sum_{k=1}^{K_i} \nabla_{\boldsymbol{w}}\ell(\boldsymbol{w}_i, \boldsymbol{z}_i(k)).$$

2.4.2. Combining one-step estimates for bounded change

It may seem natural to combine the one-step estimates using

$$\hat{\rho}_n = \max\{\tilde{\rho}_2, ..., \tilde{\rho}_n\}.$$

This method has a serious drawback. Because $\{\tilde{\rho}_i\}$ are random variables, if we combine them by taking their maximum, any particular one-step estimate $\tilde{\rho}_i$ that is large will pull up the overall estimate $\hat{\rho}_n$. This would drive $\hat{\rho}_n \to \text{diam}(\mathcal{X})$, as $n \to \infty$, resulting in a $\hat{\rho}_n$ that is trivial in the limit of large n.

We introduce an estimate from Wilson et al. (2018) that overcomes this defect. We need the following assumptions:

B.1 We have estimates $\hat{h}_W : \mathbb{R}^W \to \mathbb{R}$ that are nondecreasing in their arguments such that

$$\mathbb{E}\left[\hat{h}_W(\rho_j, ..., \rho_{j-W+1})\right] \ge \rho,$$

where $\rho_i \triangleq ||\boldsymbol{w}_i^* - \boldsymbol{w}_{i-1}^*||$.

B.2 There exists absolute constants $\{b_i\}_{i=1}^W$ for any fixed W such that $\forall \boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}_{\ge 0}^W$,

$$|\hat{h}_W(p_1, ..., p_W) - \hat{h}_W(q_1, ..., q_W)| \le \sum_{i=1}^W b_i|p_i - q_i|.$$

For example, if $\rho_i \overset{\text{iid}}{\sim} \text{Unif}[0, \rho]$, then

$$\hat{h}_W(\rho_i, \rho_{i+1}, ..., \rho_{i+W-1}) = \frac{W+1}{W}\max\{\rho_i, \rho_{i+1}, ..., \rho_{i+W-1}\}$$

is an estimator of ρ satisfied the required assumptions.

Given an estimator satisfying all of the assumptions, we let

$$\tilde{\rho}^{(i)} \triangleq \hat{h}_W(\tilde{\rho}_i, \tilde{\rho}_{i-1}, ..., \tilde{\rho}_{i-W+1})$$

and set

$$\hat{\rho}_n = \frac{1}{n-1}\sum_{i=2}^{n} \tilde{\rho}^{(i)} = \frac{1}{n-1}\sum_{i=2}^{n} \hat{h}_{\min\{W,i-1\}}(\tilde{\rho}_i, \tilde{\rho}_{i-1}, \ldots, \tilde{\rho}_{\max\{i-W+1,2\}}). \tag{2.10}$$

As shown in Wilson et al. (2018, theorem 2), the above estimator $\hat{\rho}_n$ eventually upper bounds ρ; that is, for appropriately chosen sequences $\{t_n\}$ and for all n large enough, $\hat{\rho}_n + t_n \geq \rho$ holds almost surely. See Wilson et al. (2018) for a detailed discussion of the expression of t_n.

2.5. Tracking analysis with change in minimizers unknown

We now present an extension of the results in Section 2.2, obtained by replacing ρ with its estimate given in Section 2.4. Our analysis depends on the following crucial assumptions:

C.1 For appropriate sequences $\{t_n\}$, for all n sufficiently large it holds that $\hat{\rho}_n + t_n \geq \rho$ almost surely.
C.2 $b(d_0, K_n)$ factors as $b(d_0, K_n) = \alpha(K_n)d_0^2 + \beta(K_n)$.

We have demonstrated that Assumption C.1 holds for the direct estimator $\hat{\rho}_n$. We start with a general result showing that for appropriate choices of K_n, we can control the excess risk.

Theorem 2.1 (Wilson et al., 2018, theorem 3). *Under Assumptions C.1–C.2, with $K_n \geq K^*$ for all n large enough, where K^* is defined in (2.8), we have*

$$\limsup_{n\to\infty} \left(\mathbb{E}[f_n(\mathbf{w}_n)] - f_n(\mathbf{w}_n^*) \right) \leq \epsilon \tag{2.11}$$

almost surely.

This theorem shows that for any choice of samples K_n such that $K_n \geq K^*$ holds, it follows that the excess risk can be controlled in the sense of (2.11).

To establish tracking analysis for the case where ρ is unknown, we can set

$$K_n = \min\left\{ K \geq 1 \;\middle|\; b\left(\left(\sqrt{\frac{2\epsilon}{m}} + (\hat{\rho}_{n-1} + t_{n-1}) \right)^2, K \right) \leq \epsilon \right\}, \quad n \geq 3. \tag{2.12}$$

This is the same form as the choice in (2.8) with $\hat{\rho}_{n-1} + t_{n-1}$ in place of ρ. Due to Assumption C.1, for all n large enough it holds that $\hat{\rho}_n + t_n \geq \rho$ almost surely. Then by the monotonicity assumption in A.1, for all n large enough we pick $K_n \geq K^*$ almost surely. We can therefore apply Theorem 2.1 and establish the excess risk guarantee.

3. Bound on ρ-estimate overshoot

Because we assume that the solution space $\mathcal{X}$ has bounded diameter, we always have the trivial bound

$$||\boldsymbol{w}_n^* - \boldsymbol{w}_{n-1}^*|| \leq \mathrm{diam}(\mathcal{X}).$$

An estimate of the change in minimizers $\hat{\rho}_n$, is only interesting if the bound is nontrivial; that is, $\hat{\rho}_n < \mathrm{diam}(\mathcal{X})$ when $\rho < \mathrm{diam}(\mathcal{X})$. In prior work (Wilson et al., 2018), we have proved that for sufficiently large n, $\hat{\rho}_n + t_n \geq \rho$ almost surely. In this section, we look at proving an upper bound on how much $\hat{\rho}_n$ can overshoot ρ to show that this estimate is nontrivial.

When we proved that $\hat{\rho}_n$ eventually upper bounds ρ, we did not use the fact that the points $\boldsymbol{w}_n$ at which we are evaluating the one-step estimates are approximate minimizers. In particular, that proof would still hold even if we selected the $\boldsymbol{w}_n$ randomly from the solution space $\mathcal{X}$ without using the samples $\{\boldsymbol{z}_n(k)\}_{k=1}^{K_n}$ at all. In contrast, controlling the overshoot depends critically on the fact that the points at which we evaluate the one-step estimates are approximate minimizers. The solution quality of the approximate minimizers measured by ϵ in (1.4) will control the size of the overshoot, as seen in Theorem 3.1.

To proceed with our analysis, suppose that the following conditions hold:

D.1 There exist constants $C(K_i)$ such that for all i, it holds that

$$\mathbb{E}[||\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i||^2 | \mathcal{F}_{i-1}] \leq C^2(K_i). \tag{3.1}$$

D.2 The loss function $f_n(\boldsymbol{w})$ has Lipschitz continuous gradients with parameter M; that is,

$$f_n(\boldsymbol{w}) \leq f_n(\tilde{\boldsymbol{w}}) + \langle \nabla f_n(\tilde{\boldsymbol{w}}), \tilde{\boldsymbol{w}} - \boldsymbol{w} \rangle + \frac{1}{2} M ||\tilde{\boldsymbol{w}} - \boldsymbol{w}||^2, \quad \forall \boldsymbol{w}, \tilde{\boldsymbol{w}} \in \mathcal{X}. \tag{3.2}$$

D.3 It holds that

$$\mathbb{E}[||\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{z}_i) - \nabla f_i(\boldsymbol{w})||^2 | \mathcal{F}_{i-1}] \leq \sigma^2, \quad \forall w \in \mathcal{X}. \tag{3.3}$$

Assumption D.1 is a bound on the difference in how far apart two independent outputs of the optimization algorithm $\boldsymbol{w}_i$ and $\tilde{\boldsymbol{w}}_i$ starting from $\boldsymbol{w}_{i-1}$ are. Due to the bound in Assumption A.4, we can always have the following choice of $C(Ki)$:

$$\begin{aligned}
\mathbb{E}[||\boldsymbol{w}_i - \tilde{\boldsymbol{w}}_i||^2 | \mathcal{F}_{i-1}] &\leq 2\mathbb{E}[||\boldsymbol{w}_i - \boldsymbol{w}_i^*||^2 | \mathcal{F}_{i-1}] + 2\mathbb{E}[||\tilde{\boldsymbol{w}}_i - \boldsymbol{w}_i^*||^2 | \mathcal{F}_{i-1}] \\
&\leq \frac{4}{m} b(\mathrm{diam}(\mathcal{X}), K_i) = C^2(K_i).
\end{aligned}$$

By a more sophisticated analysis, specific to the particular chosen optimization algorithm, it is possible to get tighter $C(Ki)$ bounds. Assumption D.2 imposes the Lipschitz gradient assumption on the loss function. For Assumption D.3, because $\mathbb{E}[\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{z}_i)] = \nabla f_i(\boldsymbol{w})$, it controls the variance of the stochastic gradients.

Theorem 3.1. *Suppose that all of the previously mentioned assumptions hold, and we suppose that:*

1. *The sequence of excess risks achieved, e_i, $i = 1, 2, \ldots$, satisfies*

$$\limsup_{n \to \infty} e_n \leq \epsilon. \tag{3.4}$$

2. *For all i large enough, we have that $K_i \geq \tilde{K}$ for a constant $\tilde{K}$.*

Then it follows that

$$\limsup_{n\to\infty} \mathbb{E}[\hat{\rho}_n] \leq \rho + \frac{4\sqrt{2}M\epsilon^{1/2}}{m^{3/2}} + G, \tag{3.5}$$

where

$$G \triangleq \frac{4M}{m} C(\tilde{K}) + \frac{2}{m}\left(\sigma\tilde{K}\right)^{1/2}. \tag{3.6}$$

Proof. First, we look at the one-step estimates. It holds that

$$\begin{aligned}
\tilde{\rho}_i - \rho_i &= ||\boldsymbol{w}_i - \boldsymbol{w}_{i-1}|| - ||\boldsymbol{w}_i^* - \boldsymbol{w}_{i-1}^*|| + \frac{1}{m}||\hat{G}_i|| + \frac{1}{m}||\hat{G}_{i-1}|| \\
&\leq |||\boldsymbol{w}_i - \boldsymbol{w}_{i-1}|| - ||\boldsymbol{w}_i^* - \boldsymbol{w}_{i-1}^*||| + \frac{1}{m}||\hat{G}_i|| + \frac{1}{m}||\hat{G}_{i-1}|| \\
&\leq \frac{1}{m}||\nabla f_i(\boldsymbol{w}_i)|| + \frac{1}{m}||\nabla f_{i-1}(\boldsymbol{w}_{i-1})|| + \frac{1}{m}||\hat{G}_i|| + \frac{1}{m}||\hat{G}_{i-1}|| \\
&\leq \frac{2}{m}||\nabla f_i(\boldsymbol{w}_i)|| + \frac{2}{m}||\nabla f_{i-1}(\boldsymbol{w}_{i-1})|| + \frac{1}{m}||\nabla f_i(\boldsymbol{w}_i) - \hat{G}_i|| + \frac{1}{m}||\nabla f_{i-1}(\boldsymbol{w}_{i-1}) - \hat{G}_{i-1}||.
\end{aligned}$$

By the Lipschitz gradient Assumption D.2, we have

$$||\nabla f_i(\boldsymbol{w})|| \leq M||\boldsymbol{w} - \boldsymbol{w}_i^*||.$$

Then it follows by strong convexity in Assumption A.2 that

$$||\boldsymbol{w} - \boldsymbol{w}_i^*|| \leq \sqrt{\frac{2}{m}(f_i(\boldsymbol{w}) - f_i(\boldsymbol{w}_i^*))},$$

and therefore we have

$$||\nabla f_i(\boldsymbol{w})|| \leq \frac{\sqrt{2}M}{\sqrt{m}}\sqrt{f_i(\boldsymbol{w}) - f_i(\boldsymbol{w}_i^*)}.$$

Because the square root is concave, by Jensen's inequality we have

$$\mathbb{E}\left[||\nabla f_i(\boldsymbol{w}_i)||\right] \leq \frac{\sqrt{2}Me_i^{1/2}}{m^{1/2}}.$$

This in turn implies that

$$\mathbb{E}[\tilde{\rho}_i] - \rho_i \leq \frac{2\sqrt{2}Me_i^{1/2}}{m^{3/2}} + \frac{2\sqrt{2}Me_{i-1}^{1/2}}{m^{3/2}} + \frac{1}{m}\mathbb{E}||\nabla f_i(\boldsymbol{w}_i) - \hat{G}_i|| + \frac{1}{m}\mathbb{E}||\nabla f_{i-1}(\boldsymbol{w}_{i-1}) - \hat{G}_{i-1}||.$$

Next, we look at bounding $\mathbb{E}||\nabla f_i(\boldsymbol{w}_i) - \hat{G}_i||$. For $\tilde{\boldsymbol{w}}_i \in \mathcal{X}$, denote

$$\tilde{G}_i \triangleq \frac{1}{K_i}\sum_{k=1}^{K_i} \nabla_{\boldsymbol{w}} \ell(\tilde{\boldsymbol{w}}_i, \boldsymbol{z}_i(k)).$$

Then we have

$$\begin{aligned}
||\nabla f_i(\boldsymbol{w}_i) - \hat{G}_i|| &\leq ||\hat{G}_i - \tilde{G}_i|| + ||\tilde{G}_i - \nabla f_i(\tilde{\boldsymbol{w}}_i)|| + ||\nabla f_i(\tilde{\boldsymbol{w}}_i) - \nabla f_i(\boldsymbol{w}_i)|| \\
&\leq ||\hat{G}_i - \tilde{G}_i|| + ||\tilde{G}_i - \nabla f_i(\tilde{\boldsymbol{w}}_i)|| + M||\tilde{\boldsymbol{w}}_i - \boldsymbol{w}_i||.
\end{aligned}$$

Using the direct estimate lower bound analysis from Wilson et al. (2018) it follows that

$$\mathbb{E}[||\nabla f_i(\boldsymbol{w}_i) - \hat{G}_i|||\mathscr{F}_{i-1}] \le MC(K_i) + \left(\frac{\sigma}{K_i}\right)^{1/2} + MC(K_i). \tag{3.7}$$

This shows that

$$\begin{aligned}\mathbb{E}[\tilde{\rho}_i] - \rho_i \le \frac{2\sqrt{2}M\epsilon_i^{1/2}}{m^{3/2}} + \frac{2\sqrt{2}M\epsilon_{i-1}^{1/2}}{m^{3/2}} + \mathbb{E}\left[\frac{2M}{m}C(K_i) + \frac{1}{m}\left(\frac{\sigma}{K_i}\right)^{1/2}\right] \\ + \mathbb{E}\left[\frac{2M}{m}C(K_{i-1}) + \frac{1}{m}\left(\frac{\sigma}{K_{i-1}}\right)^{1/2}\right].\end{aligned} \tag{3.8}$$

Then, plugging in the definition of $\hat{\rho}_n$ it follows that

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}[\hat{\rho}_n] &\le \rho + \frac{4\sqrt{2}M\epsilon^{1/2}}{m^{3/2}} + \frac{4M}{m}C(\tilde{K}) + \frac{2}{m}(\sigma\tilde{K})^{1/2} \\ &= \rho + \frac{4\sqrt{2}M\epsilon^{1/2}}{m^{3/2}} + G.\end{aligned} \tag{3.9}$$

□

This shows that the direct estimate is a nontrivial upper bound for sufficiently small ϵ. Note that, in practice, the $\tilde{K}$ will be a function of ϵ, because we can pick $\tilde{K} = K^*$ with K^* defined in (2.8). Note that K^* is itself a function of ϵ. This means that the G term in (3.9), which is a function of $\tilde{K}$ is also a function of ϵ. Thus, the entire overshoot term is a function of ϵ and, in fact, by inspection, it goes to zero as $\epsilon \to 0$ if $K^* \to \infty$ as $\epsilon \to 0$ (as K^* defined in 2.8 does).

4. Extensions relevant to machine learning applications

4.1. Cost approach

A natural way to assess the usefulness of our approach is to choose a number of samples $\{K_n\}_{n=1}^T$ over a horizon of length T using the choice in (2.12), and compare against taking $\sum_{n=1}^T K_n$ samples at time $n = 1$ and no samples at the other T – 1 time instants. See Section 5 for such a comparison.

In this section, we consider a different type of comparison based on assuming that there is a cost $p(K_n)$ of taking K_n samples. For example, we could have

$$p(K) = P_0 \mathbb{1}_{\{K>0\}} + P_1 K. \tag{4.1}$$

This implies that we pay a fixed cost of P_0 any time we take at least one sample and a marginal cost of P_1 per sample. We want to control the excess risk by deciding when to take samples, and how many samples to take with a total budget P over a horizon of length T; that is,

$$\sum_{n=1}^{T} p(K_n) \le P. \tag{4.2}$$

For the option of taking all samples up front:

$$K_n = \begin{cases} \max\{K \geq 1 | p(K) \leq P\}, & n = 1 \\ 0, & 2 \leq n \leq T. \end{cases} \tag{4.3}$$

Another option is to sample every ΔT time instants and divide the cost budget evenly over the times that we take samples using

$$K_n = \begin{cases} \max\left\{K \geq 1 | p(K) \leq \left\lfloor \frac{P}{T/\Delta T} \right\rfloor \right\}, & \text{if } \Delta T \text{ divides } (n-1) \\ 0, & \text{else.} \end{cases} \tag{4.4}$$

For analysis, we need Assumption C.1 and the following additional assumptions:

E.1 There exists a function $b'(||\boldsymbol{w} - \boldsymbol{w}_n^*||^2)$ such that

$$f_n(\boldsymbol{w}) - f_n(\boldsymbol{w}_n^*) \leq b'(||\boldsymbol{w} - \boldsymbol{w}_n^*||^2). \tag{4.5}$$

For example, suppose that the functions $f_n(\boldsymbol{w})$ have Lipschitz continuous gradients with modulus M and $\boldsymbol{w}_n^* \in \text{int}(\mathcal{X})$ for all $n \geq 1$, where $\text{int}(\mathcal{X})$ is the interior of $\mathcal{X}$. By the descent lemma given in Bertsekas (1999), we have

$$\begin{aligned} f_n(\boldsymbol{w}_n) - f_n(\boldsymbol{w}_n^*) &\leq \langle \nabla f_n(\boldsymbol{w}_n^*), \boldsymbol{w}_n - \boldsymbol{w}_n^* \rangle + \frac{1}{2} M ||\boldsymbol{w}_n - \boldsymbol{w}_n^*||^2 \\ &= \frac{1}{2} M ||\boldsymbol{w}_n - \boldsymbol{w}_n^*||^2. \end{aligned}$$

Thus, we can set

$$b'(||\boldsymbol{w}_n - \boldsymbol{w}_n^*||) = \frac{1}{2} M ||\boldsymbol{w}_n - \boldsymbol{w}_n^*||^2.$$

Because we need to consider the possibility that $K_n = 0$ for some n in $\{1, \ldots, T\}$ but still provide estimates of the excess risk, we need an alternate version of the bound in (2.5). Define

$$t_s(n) \triangleq \max\{m | 1 \leq m \leq n \text{and} K_m > 0\}, \tag{4.6}$$

where $t_s(n)$ is the last time no later than n at which samples were taken. If no samples have been taken so far, then by convention $t_s(n) = +\infty$. We construct the recursively defined function $\tilde{b}_n(\rho, K_n)$ by considering the following four cases:

1. No samples have been taken by time n:

$$\tilde{b}_n(\rho, K_n) \triangleq b'(\text{diam}(\mathcal{X})).$$

2. Samples taken at time n for the first time:

$$\tilde{b}_n(\rho, K_n) \triangleq b(\text{diam}(\mathcal{X}), K_n).$$

3. No samples taken at time n but samples have been taken previously:

$$\tilde{b}_n(\rho, K_n) \triangleq e\left(\sqrt{\frac{2}{m} \tilde{b}_{t_s(n-1)}} + \left((n - t_s(n-1))\rho\right)\right).$$

4. Samples taken at time n and samples have been taken previously:

$$\tilde{b}_n(\rho, K_n) \triangleq b\left(\sqrt{\frac{4}{m}\tilde{b}_{t_s(n-1)} + 2\Big((n - t_s(n-1))\rho\Big)^2}, K_n\right),$$

where $\tilde{b}_{t_s(n-1)}$ is the bound on the excess risk at time $t_s(n-1)$.

Suppose that over a time horizon of length T we have a total cost budget P with respect to the number of samples $\{K_n\}_{n=1}^T$ as in (4.2). Define the *excess risk gaps*

$$\xi_n \triangleq \left(\tilde{b}_n(\rho, K_n) - \epsilon\right)_+ \tag{4.7}$$

with $(x)_+ = \max\{x, 0\}$. The variable ξ_n is the extent to which the target excess risk of ϵ is violated upwards. If our excess risk is below our target level ϵ, then we set $\xi_n = 0$. Our goal is to minimize the size of the ξ_n, while taking into account the cost constraint in (4.2). To control the size of ξ_n, suppose that we have a function $\phi : \mathbb{R}^T \to \mathbb{R}$ that describes the cumulative loss of the excess risk gaps $\xi_1, \ldots, \xi_T$.

We now provide some possible choices for $\phi(\xi_1, \ldots, \xi_T)$:

$$\phi(\xi_1, \ldots, \xi_T) = \frac{1}{T}\sum_{n=1}^{T} \xi_t, \tag{4.8}$$

$$\phi(\xi_1, \ldots, \xi_T) = \max\{\xi_1, \ldots, \xi_T\}, \tag{4.9}$$

$$\phi(\xi_1, \ldots, \xi_T) = \max_{(a,b)\in\tau} \sum_{n=a}^{b} \xi_n, \tag{4.10}$$

with

$$\tau = \{(a,b) | a < b, \xi_a \le \xi_{a+1} \le \cdots \le \xi_b\}.$$

The choices given in (4.8) and (4.9) penalize the average and maximum excess risk gaps respectively. In practice, with these choices, we will stop taking samples before the horizon T resulting in relatively poor performance toward the end of the horizon. The third choice gets around this problem by penalizing large increasing runs of excess risk gaps, and tends to favor a more uniform choice of the number of samples K_n.

We first consider the case when ρ is known to us and plan over the horizon of length T by solving the following optimization problem:

$$\begin{array}{lll} \underset{K_1, \ldots, K_T}{\text{minimize}} & \phi(\xi_1, \ldots, \xi_T) & \\ \text{subject to} & \displaystyle\sum_{n=1}^{T} p(\rho, K_n) \le P & \\ & \mathbb{1}_{\{K_1 > 0\}} \le \mathbb{1}_{\{K_2 > 0\}} & \\ & \mathbb{1}_{\{K_n > 0\}} \le \mathbb{1}_{\{K_{n-1} > 0\}} + \mathbb{1}_{\{K_{n+1} > 0\}} & n = 2, \ldots, T-1 \\ & \mathbb{1}_{\{K_{T-1} > 0\}} \le \mathbb{1}_{\{K_T > 0\}} & \\ & K_n \in \mathbb{Z}_{\ge 0} & n = 1, \ldots, T. \end{array} \tag{4.11}$$

The idea of this problem is to satisfy the excess risk bound ϵ with minimal violation $\phi(\xi_1, \ldots, \xi_T)$.

To estimate ρ, we need samples from consecutive time instants. Therefore, we impose the constraint that if we take samples at time n, then we must take samples at either time $n - 1$ or time $n + 1$ through the constraint

$$\mathbb{1}_{\{K_n>0\}} \leq \mathbb{1}_{\{K_{n-1}>0\}} + \mathbb{1}_{\{K_{n+1}>0\}}.$$

The problem in (4.11) is a mixed integer non-linear programming problem (MINLP). There are no general methods to efficiently solve this MINLP, and we therefore consider a relaxation of this problem later.

In the case that we know ρ, we can plan the number of samples ahead of time before any samples have been taken. When ρ is unknown, we cannot plan over the entire horizon. Instead, at each time instant m we have to plan over the remaining time horizon of length $T - m + 1,$ while using the estimate $\hat{\rho}_{m-1} + t_{m-1}$ in place of ρ and the remaining cost budget

$$P - \sum_{n=1}^{m-1} p(K_n).$$

We then consider the cost-to-go problem

$$\begin{aligned} \underset{K_m,\dots,K_T}{\text{minimize}} \quad & \phi(\xi_m, \dots, \xi_T) \\ \text{subject to} \quad & \sum_{n=m}^{T} p(K_n) \leq P - \sum_{n=1}^{m-1} p(K_n) \\ & \mathbb{1}_{\{K_m>0\}} \leq \mathbb{1}_{\{K_{m+1}>0\}} \\ & \mathbb{1}_{\{K_m>0\}} \leq \mathbb{1}_{\{K_{m-1}>0\}} + \mathbb{1}_{\{K_{m+1}>0\}} \qquad n = m+1, \dots, T-1 \\ & \mathbb{1}_{\{K_{T-1}>0\}} \leq \mathbb{1}_{\{K_T>0\}} \\ & K_n \in \mathbb{Z}_{\geq 0} \qquad n = m, \dots, T. \end{aligned} \tag{4.12}$$

This is the same form as (4.11), except that it is over the time horizon from $n = m, \dots, T,$ taking into account the portion of the cost budget that has been expended. In this problem, we only optimize over $K_m, \dots, K_T$. This problem is again an MINLP.

Next, we look at approximate solutions to (4.11) and (4.12). The major difficulties in solving these programs are that the decision variables $\{K_n\}_{n=1}^{T}$ are integer valued and the cost function $p(K)$ may be discontinuous at zero due to fixed costs. We consider relaxing K_n to be real valued and introduce a piecewise approximation $\hat{p}(K)$ of the cost functions $p(K)$:

$$\hat{p}(K) = \left(\frac{p(K_0)K}{K_0}\right) \mathbb{1}_{\{K \leq K_0\}} + p(K)\ \mathbb{1}_{\{K>K_0\}}.$$

Generally, we pick $0 < K_0 < 1$. We consider the relaxed program

$$\begin{aligned} \underset{K_1,\dots,K_T}{\text{minimize}} \quad & \phi(\xi_1, \dots, \xi_T) \\ \text{subject to} \quad & \sum_{n=1}^{T} \hat{p}(\rho, K_n) \leq P \\ & K_1 \leq K_2 \\ & K_n \leq K_{n-1} + K_{n+1} \qquad n = 2, \dots, T-1 \\ & K_{T-1} \leq K_T \\ & K_n \in \mathbb{R}_{\geq 0} \qquad n = 1, \dots, T. \end{aligned} \tag{4.13}$$

We also relax the indicator constraints to inequality to encourage taking samples at consecutive times. In practice, this forces more gradual changes in samples K_n and makes it easier to solve these problems. This problem can be readily solved by gradient-based solvers such as the Interior Point Optimizer (IPOPT) in Wächter and Biegler (2006).

When ρ is unknown, we can repeatedly solve this problem using the latest estimate of ρ by solving the following sequence of problems:

$$\begin{aligned} \underset{K_{n+1},\ldots,K_T}{\text{minimize}} \quad & \phi(\xi_1,\ldots,\xi_T) \\ \text{subject to} \quad & \sum_{i=1}^{n} \hat{p}(\hat{\rho}_i, K_i) \le P - \sum_{i=n+1}^{T} \hat{p}(\hat{\rho}_{i-1}, K_i) \\ & K_1 \le K_2 \\ & K_n \le K_{n-1} + K_{n+1} && n = 2,\ldots,T-1 \\ & K_{T-1} \le K_T \\ & K_n \in \mathbb{R}_{\ge 0} && n = 1,\ldots,T. \end{aligned} \tag{4.14}$$

4.2. Cross-validation

We can also apply cross-validation for model selection. Suppose we have loss functions $\ell_\lambda(\boldsymbol{w},\boldsymbol{z})$ parameterized by λ, which controls the model complexity. For example, we could have a quadratic penalty term

$$\ell_\lambda(\boldsymbol{w},\boldsymbol{z}) = \tilde{\ell}(\boldsymbol{w},\boldsymbol{z}) + \frac{1}{2}\lambda||\boldsymbol{w}||^2. \tag{4.15}$$

The value of $\lambda = 0$ corresponds to the true loss function that we want to minimize. Suppose that we have C different values $\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(C)}$ of λ under consideration. For each $\lambda^{(i)}$, we generate an approximate minimizer $\boldsymbol{w}_n^{(i)}$ of

$$\mathbb{E}_{\boldsymbol{z}_n \sim p_n}[\ell_{\lambda^{(i)}}(\boldsymbol{w},\boldsymbol{z}_n)]. \tag{4.16}$$

We want to select the value $\lambda^{(i)}$ and corresponding $\boldsymbol{w}_n^{(i)}$ that achieves the smallest loss

$$\mathbb{E}_{\boldsymbol{z}_n \sim p_n}[\ell_0(\boldsymbol{w}_n^{(i)},\boldsymbol{z}_n)]. \tag{4.17}$$

We generate an approximate minimizer $\boldsymbol{w}_n^{(i)}$ for each problem in (4.16) starting from $\boldsymbol{w}_{n-1}^{(i)}$. To select the best choice of $\lambda^{(i^*)}$ in terms of minimizing (4.17), we apply cross-validation and set $\boldsymbol{w}_n = \boldsymbol{w}_n^{(i^*)}$ (see Hastie et al., 2001).

The idea behind cross-validation is to divide the training samples $\{\boldsymbol{z}_n(k)]\}_{k=1}^{K_n}$ into P equal-sized pieces. For every P – 1 out of P pieces, we use the P – 1 pieces of the training set to generate an approximate solution $\tilde{\boldsymbol{w}}_n^{(i)}$ to (4.16). We use the remaining piece of the training set to evaluate the empirical test loss achieved by $\tilde{\boldsymbol{w}}_n^{(i)}$ using a sample average approximation. We do this for every possible choice of P – 1 out of P pieces and average the empirical test loss estimates. We then select the value $\lambda^{(i^*)}$ that achieves the smallest empirical test loss.

To apply cross-validation to our framework, we run C parallel versions of our approach and at time n we generate C different choices for the number of samples $K_n^{(i)}$. We then choose

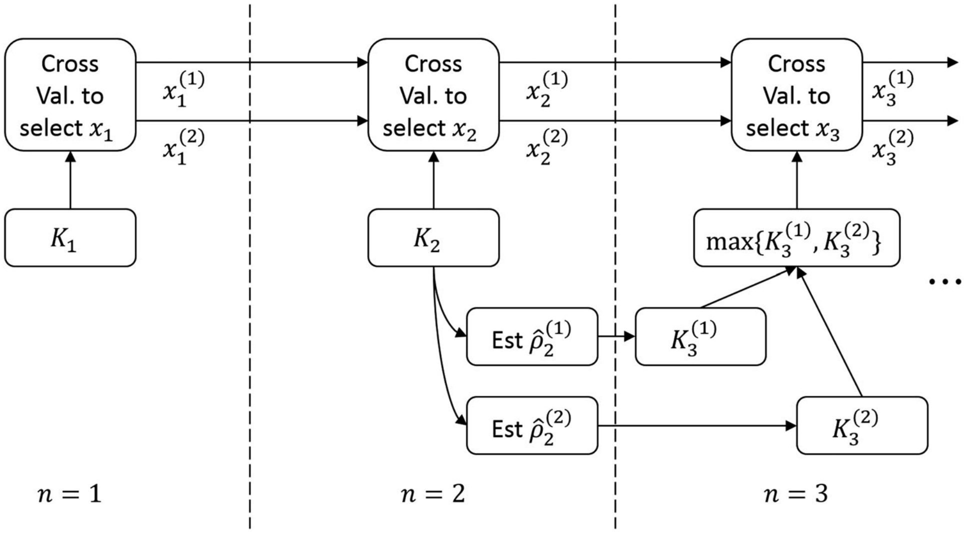

Figure 1. Cross-validation approach.

$$K_n = \max\{K_n^{(1)}, \ldots, K_n^{(C)}\}.$$

After choosing K_n, we apply the usual cross-validation approach to select $\lambda^{(i)}$ for time n. Figure 1 shows this approach for two values of λ.

5. Experiments

We provide two regression examples for synthetic and real data as well as a classification example for synthetic data. For the synthetic regression problem, we can explicitly compute ρ and $\boldsymbol{w}_n^*$ and exactly evaluate the performance of our method. It is straightforward to check that all requirements in A.1–A.5 are satisfied for the problems considered in this section. We apply the “do not update past excess risk” choice of K_n here.

5.1. Synthetic regression

Consider a regression problem with synthetic data using the penalized quadratic loss

$$\ell(\boldsymbol{w}, \boldsymbol{z}) = \frac{1}{2}(y - \boldsymbol{x}^\top \boldsymbol{w})^2 + \frac{1}{2}\lambda||\boldsymbol{w}||^2$$

with $\boldsymbol{z} = (\boldsymbol{x}, y) \in \mathbb{R}^3$. We further assume that

$$\boldsymbol{z}_n \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_{\boldsymbol{x}}^2 \boldsymbol{I} & r_{\boldsymbol{x}_n, y_n} \\ r_{\boldsymbol{x}_n, y_n}^\top & \sigma_{y_n}^2 \end{bmatrix}\right).$$

Under these assumptions, we can analytically compute minimizers $\boldsymbol{w}_n^*$ of $f_n(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{z}_n \sim p_n}[\ell(\boldsymbol{w}, \boldsymbol{z}_n)]$. We change only $r_{\boldsymbol{x}_n, y_n}$ and $\sigma_{y_n}^2$ appropriately to ensure that $||\boldsymbol{w}_n^* - \boldsymbol{w}_{n-1}^*|| = \rho$ holds for all n. We find approximate minimizers using SGD with $\lambda = 0$. We estimate ρ using the direct estimate.

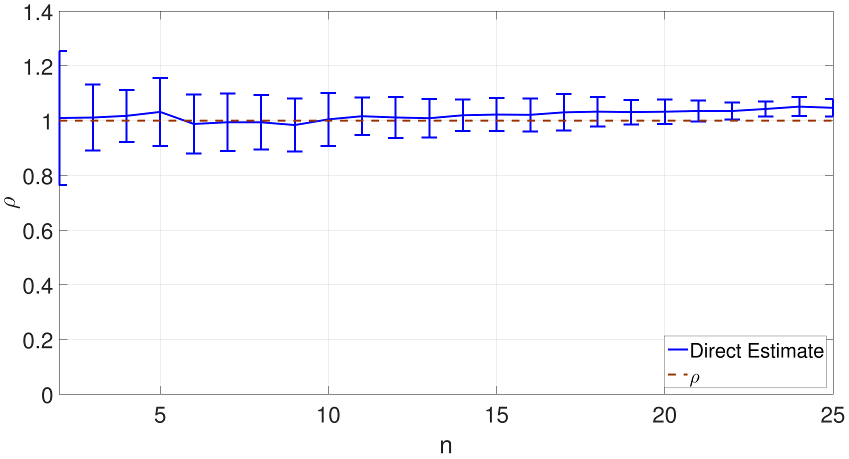

Figure 2. ρ Estimate for synthetic regression.

We let n range from 1 to 25 with $\rho = 1$, a target excess risk $\epsilon = 0.1$, and K_n from (2.12). We average over 20 runs of our algorithm. Figure 2 shows $\hat{\rho}_n$, our estimate of ρ, which is above ρ in general. Figure 3 shows the number of samples K_n, which settles down as the estimate of ρ converges. We can exactly compute $f_n(\boldsymbol{w}_n) - f_n(\boldsymbol{w}_n^*)$ and thus, by averaging over the 20 runs of our algorithm, we can estimate the excess risk (denoted "sample average estimate"). We cover the time horizon from $n = 1$ to 25 to yield the sample average estimate excess risk given by $2.797 \times 10^{-2} \pm 1.071 \times 10^{-2}$. Therefore, we see that we achieve our desired excess risk.

5.1.1. Cost approach

We consider applying the cost approach in Section 4.1 to the synthetic regression problem with the cost in (4.1). We compare the optimal cost approach introduced in (4.13) of Section 4.1 to the approach in (2.12), taking all samples at time $n = 1$ as in (4.3) and taking samples every five time instants as in (4.4). Note that the method from (2.12) does not satisfy the cost budget. Figure 4 shows the test loss of these approaches. We achieve test loss similar to the method in (2.12) and better than the other two methods. Figure 5 shows the number of samples selected for both methods. At some time instants, our optimal cost approach does not take samples.

5.2. Synthetic classification

Consider a binary classification problem using

$$\ell(\boldsymbol{w}, \boldsymbol{z}) = \frac{1}{2}(1 - y(\boldsymbol{x}^\top \boldsymbol{w}))_+^2 + \frac{1}{2}\lambda||\boldsymbol{w}||^2$$

with $\boldsymbol{z} = (\boldsymbol{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ and $(y)_+ = \max\{y, 0\}$. This is a smoothed version of the hinge loss used in support vector machines Hastie et al., 2001). We suppose that at time n, the two classes have features drawn from a Gaussian distribution with covariance matrix $\sigma^2 \boldsymbol{I}$ but different means $\mu_n^{(1)}$ and $\mu_n^{(2)}$; that is,

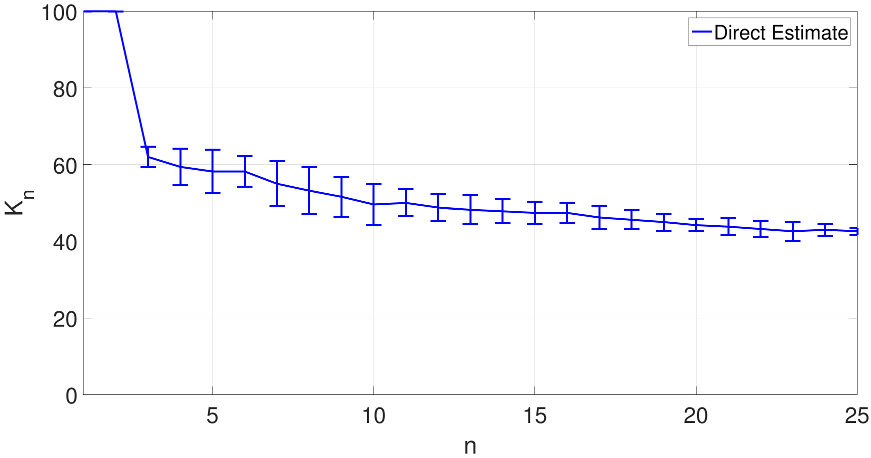

Figure 3. K_n for synthetic regression.

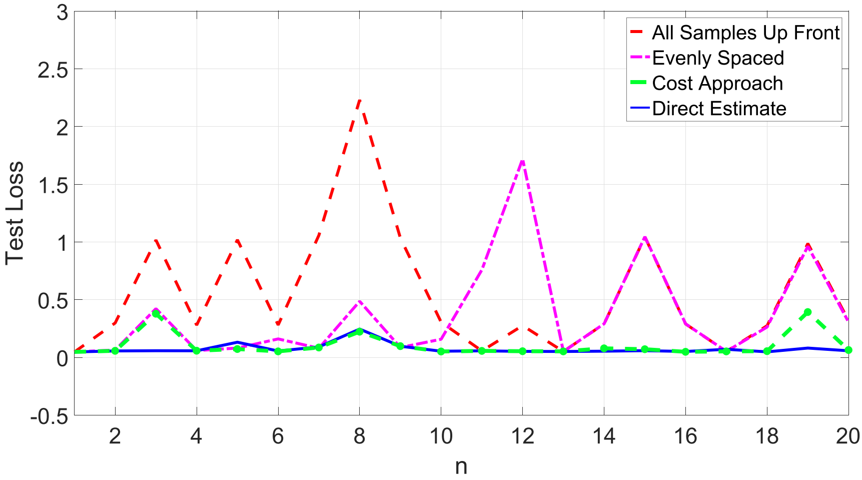

Figure 4. Test loss for synthetic regression with cost approach.

$$\boldsymbol{x}_n | \{y_n = i\} \sim \mathcal{N}(\mu_n^{(i)}, \sigma^2 \boldsymbol{I}).$$

The class means move slowly over uniformly spaced points on a unit sphere in $\mathbb{R}^d$ as in Figure 6 to ensure that the constant Euclidean norm condition $||\boldsymbol{w}_n^* - \boldsymbol{w}_{n-1}^*|| = \rho$ holds. We find approximate minimizers using SGD with $\lambda = 0.1$. We estimate ρ using the direct estimate.

We let n range from 1 to 25 and target an excess risk $\epsilon = 0.1$. We average over 20 runs of our algorithm. As a comparison, if our algorithm takes $\{K_n\}_{n=1}^{25}$ samples, then we consider taking $\sum_{n=1}^{25} K_n$ samples up front at $n = 1$. This is what we would do if we assumed that our problem is not time varying. Figure 7 shows $\hat{\rho}_n$, our estimate of ρ. Figure 8 shows the average test loss for both sampling strategies. To compute the test loss we draw T_n additional samples $\{\boldsymbol{z}_n^{\text{test}}(k)\}_{k=1}^{T_n}$ from p_n and compute $\frac{1}{T_n}\sum_{k=1}^{T_n} \ell(\boldsymbol{w}_n, \boldsymbol{z}_n^{\text{test}}(k))$. We see that our approach achieves substantially smaller test loss than taking all samples up front. We do not draw the error bars on this plot because it makes it difficult to see the actual losses achieved.

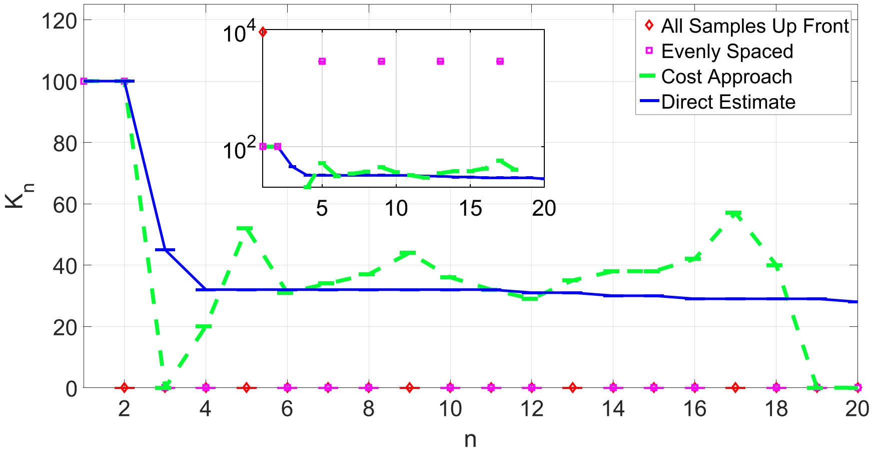

Figure 5. K_n for synthetic regression with cost approach.

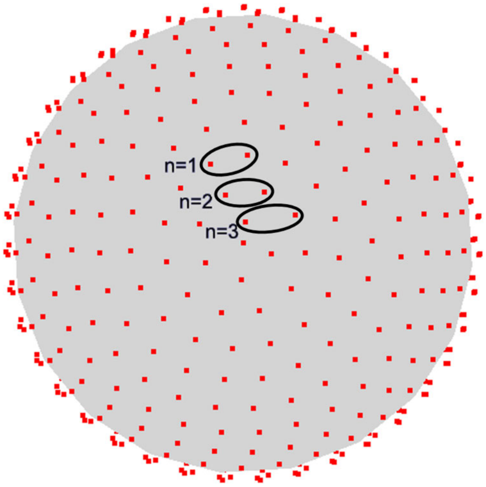

Figure 6. Evolution of class means.

To further evaluate our approach, we look at the receiver operating characteristic (ROC) of our classifiers. The ROC is a plot of the probability of a true positive against the probability of a false positive. The area under the curve (AUC) of the ROC equals the probability that a randomly chosen positive instance ($y = 1$) will be rated higher than a negative instance ($y = -1$) Fawcett, 2006). Thus, a large AUC is desirable. Figure 9 plots the AUC of our approach against taking all samples up front. Our sampling approach achieve a substantially larger AUC.

5.3. Panel study on income dynamics–regression

The Panel Study of Income Dynamics surveyed individuals every year to gather demographic and income data annually from 1974 to 2012 (Brown et al., 2015). We want to predict an individual's annual income (y) from several demographic features ($\boldsymbol{x}$),

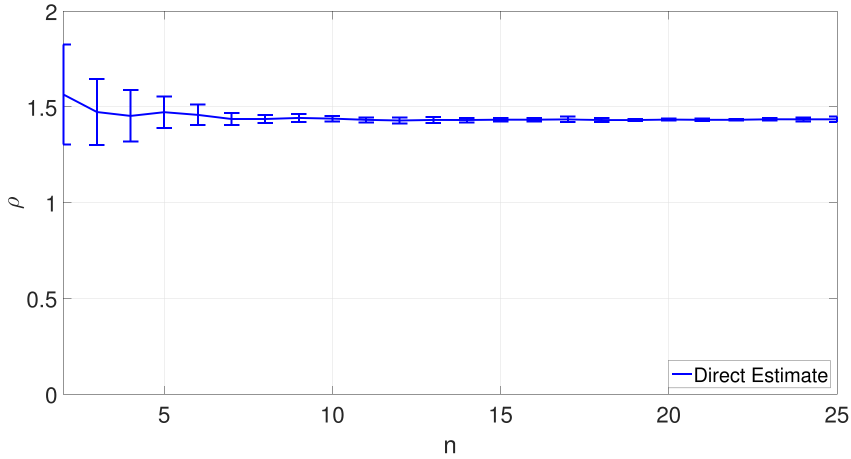

Figure 7. Estimate of ρ for synthetic classification.

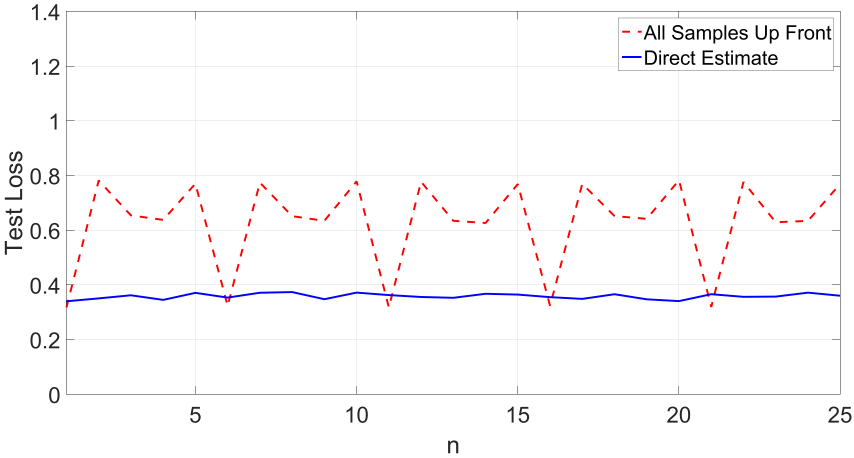

Figure 8. Test loss for synthetic classification.

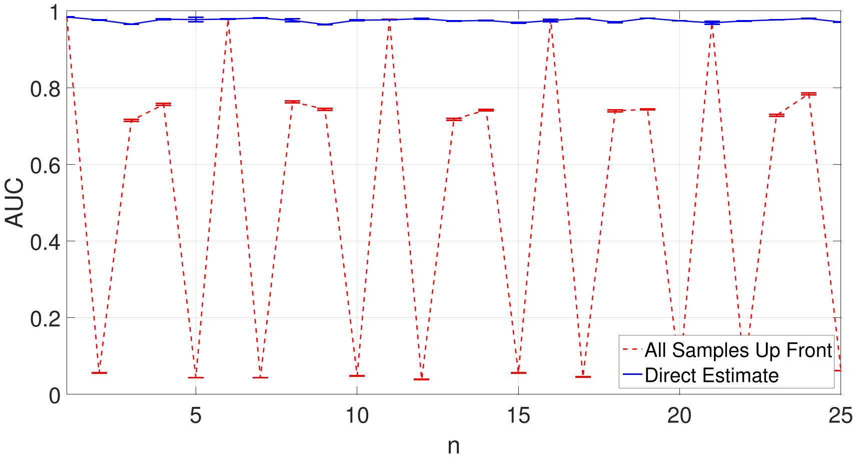

Figure 9. Area under the curve for synthetic classification.

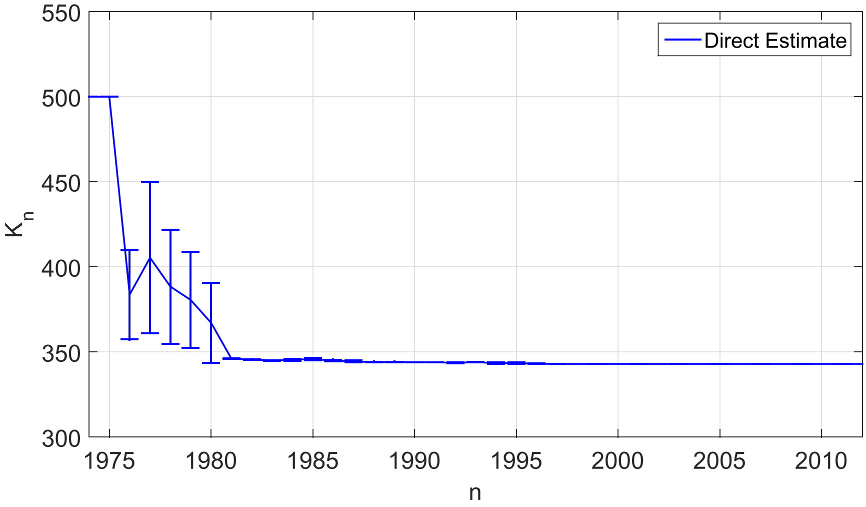

Figure 10. Number of samples K_n in Panel Study on Income Dynamics regression problem.

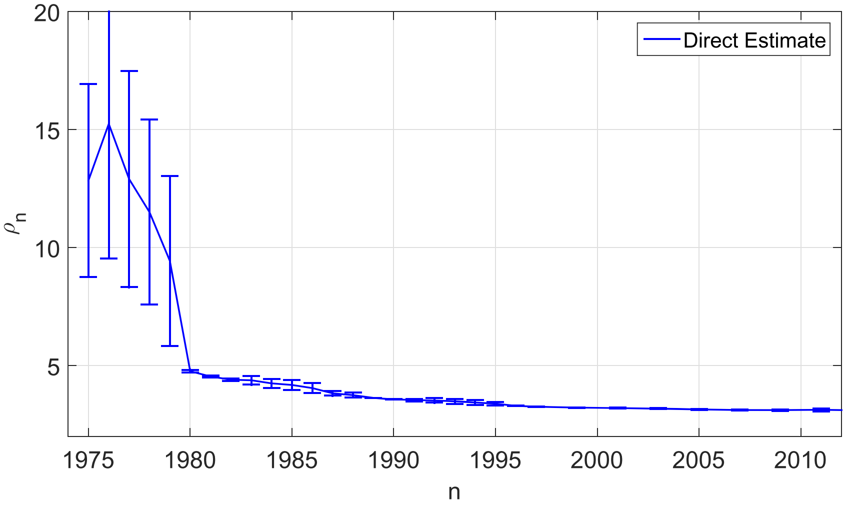

Figure 11. Estimate of ρ in Panel Study on Income Dynamics regression problem.

including age, education, work experience, etc., chosen based on previous economic studies in Murphy and Welch (1990).

The idea of this problem conceptually is to rerun the survey process and determine how many samples we would need if we wanted to solve this regression problem to within a desired excess risk criterion ϵ.

We use the same loss function, direct estimate for ρ, and minimization algorithm as the synthetic regression problem. We average over 20 runs of our algorithm by resampling without replacement (see Hastie et al., 2001). For the sake of comparison, given a choice of samples $\{K_n\}_{n=1}^T$ produced by our approach, we compare against taking $\sum_{n=1}^T K_n$ samples at time $n = 1$ and none afterwards. Note that this is what we would do if we believed that the regression model does not change over time. We are aware of

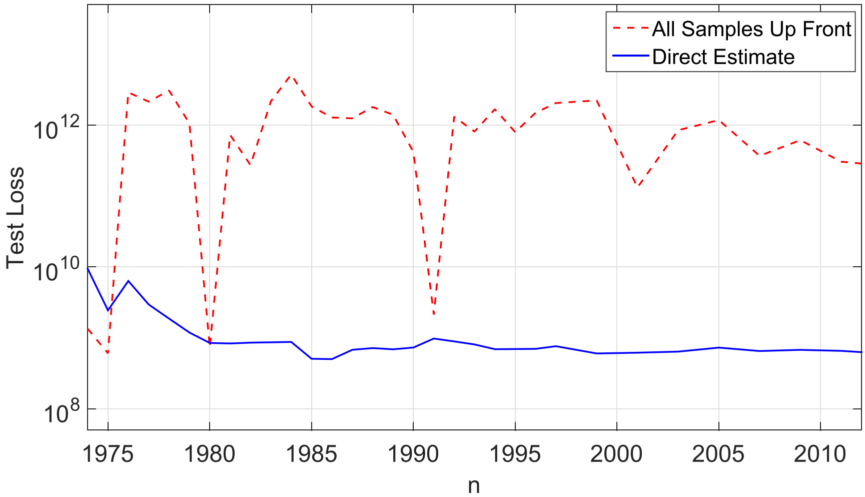

Figure 12. Test loss in Panel Study on Income Dynamics regression problem.

no other methods to select the number of samples K_n to control the excess risk against which we could compare our approach.

Figure 10 shows the number of samples K_n, which settles down quickly. Figure 11 shows $\hat{\rho}_n$. Figure 12 shows the test losses over time evaluated over 20% of the available samples. The test loss for our approach is substantially less than that obtained by taking the same number of samples up front.

6. Conclusion

We introduced a framework for adaptively solving a sequence of learning problems. We developed estimates of the change in the minimizers used to determine the number of training samples K_n needed to achieve a target excess risk ϵ. We introduced a cost-based approach to select the number of samples and an approach to apply cross-validation. Experiments with synthetic and real data demonstrate that this approach is effective.

Funding

This work was supported in part by the U.S. National Science Foundation under awards CCF 1111342 and NSF DMS 1312907, and in part by the U.S. Army Research Laboratory under cooperative agreement W911NF-17-2-0196, through the University of Illinois at Urbana–Champaign.

References

Agarwal, A., Daumé, H., and Gerber, S. (2010). Learning Multiple Tasks Using Manifold Regularization, in *Proceedings of Advances in Neural Information Processing Systems*, pp. 46–54.

Agarwal, A., Rakhlin, A., and Bartlett, P. (2008). Matrix Regularization Techniques for Online Multitask Learning, Technical Report UCB/EECS-2008-138, EECS Department, University of California, Berkeley.

Bartlett, P., Hazan, E., and Rakhlin, A. (2008). Adaptive Online Gradient Descent, in Proceedings of *Advances in Neural Information Processing Systems*, pp. 65–72.

Bengio, Y. and Frasconi, P. (1996). Input–Output HMM's for Sequence Processing, *IEEE Transactions on Neural Networks* 7: 1231–1249. doi:10.1109/72.536317

Bertsekas, D. (1999). *Nonlinear Programming*, Boston: Athena Scientific.

Brown, C., Freedman, V., Sastry, N., McGonagle, K., Pfeffer, F., Schoeni, R., and Stafford, F. (2015). *Panel Study of Income Dynamics, Public Use Dataset*, Ann Arbor: University of Michigan.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*, Cambridge: Cambridge University Press.

Chiang, C., Yang, T., Lee, C., Mahdavi, M., Lu, C., Jin, R., and Zhu, S. (2012). Online Optimization with Gradual Variations, in *Proceedings of Conference on Learning Theory*, vol. 23, pp. 6.1–6.20.

Dietterich, T. (2002). Machine Learning for Sequential Data: A Review, in *Structural, Syntactic, and Statistical Pattern Recognition*, pp. 15–30.

Dontchev, A. and Rockafellar, R. (2009). *Implicit Functions and Solution Mappings: A View from Variational Analysis*, New York: Springer.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *Journal of Machine Learning Research* 12: 2121–2159.

Duchi, J. and Singer, Y. (2009). Efficient Online and Batch Learning Using Forward Backward Splitting, *Journal of Machine Learning Research* 10: 2899–2934.

Evgeniou, T. and Pontil, M. (2004). Regularized Multi-Task Learning, in *Proceedings of Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 109–117.

Fawcett, T. (2006). An Introduction to ROC Analysis, *Pattern Recognition Letters* 27: 861–874. doi:10.1016/j.patrec.2005.10.010

Fawcett, T. and Provost, F. (1997) Adaptive Fraud Detection, *Data Mining and Knowledge Discovery* 1: 291–316. doi:10.1023/A:1009700419189

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic Regret Algorithms for Online Convex Optimization, *Machine Learning* 69: 169–192. doi:10.1007/s10994-007-5016-8

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*, Boston: MIT Press.

Murphy, K. and Welch, F. (1990). Empirical Age-Earnings Profiles, *Journal of Labor Economics*, 8: 202–229. doi:10.1086/298220

Pan, S. and Yang, Q. (2010). A Survey on Transfer Learning, *IEEE Transactions on Knowledge and Data Engineering* 22: 1345–1359. doi:10.1109/TKDE.2009.191

Qian, N. and Sejnowski, T. (1988). Predicting the Secondary Structure of Globular Proteins Using Neural Network Models, *Journal of Molecular Biology* 202: 865–884. doi:10.1016/0022-2836(88)90564-5

Rakhlin, A. and Sridharan, K. (2012). Online Learning with Predictable Sequences, *arXiv: 1208*.3728.

Shalev-Shwartz, S. and Kakade, S. (2009). Mind the Duality Gap: Logarithmic Regret Algorithms for Online Optimization, in *Proceedings of Advances in Neural Information Processing Systems*, pp. 1457–1464.

Shalev-Shwartz, S. and Singer, Y. (2007). Convex Repeated Games and Fenchel Duality, in *Proceedings of Advances in Neural Information Processing Systems*, pp. 1265–1271.

Shalev-Shwartz, S. and Singer, Y. (2007). Logarithmic Regret Algorithms for Strongly Convex Repeated Games, Technical Report, Hebrew University.

Towfic, Z., Chu, J., and Sayed, A. (2013). Online Distributed Online Classification in the Midst of Concept Drifts, *Neurocomputing* 112: 138–152. doi:10.1016/j.neucom.2012.12.043

Wächter, A. and Biegler, L. T. (2006). On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming, *Mathematical Programming* 106: 25–57. doi:10.1007/s10107-004-0559-y

Wilson, C. and Veeravalli, V. (2016a). Adaptive Sequential Learning, in *Proceedings of Asilomar Conference on Signals, Systems and Computers*, pp. 326–330.

Wilson, C. and Veeravalli, V. (2016b). Adaptive Sequential Optimization with Applications to Machine Learning, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2642–2646.

Wilson, C., Veeravalli, V. V., and Nedić, A. (2018). Adaptive Sequential Stochastic Optimization, *IEEE Transactions on Automatic Control* 64: 496–509. doi:10.1109/TAC.2018.2816168

Xiao, L. (2010). Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization, *Journal of Machine Learning Research* 11: 2543–2596.

Zhang, Y. and Yeung, D. (2012). A Convex Formulation for Learning Task Relationships in Multi-Task Learning, *arXiv: 1203*.3536.

Zinkevich, M. (2003). Online Convex Programming and Generalized Infinitesimal Gradient Ascent, in *Proceedings of International Conference on Machine Learning*, pp. 928–936.