ADAPTIVE SEQUENTIAL OPTIMIZATION WITH APPLICATIONS TO MACHINE LEARNING

Craig Wilson and Venugopal V. Veeravalli

Coordinated Science Lab and Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
{wilson60,vvv}@illinois.edu

ABSTRACT

A framework is introduced for solving a sequence of slowly changing optimization problems, including those arising in regression and classification applications, using optimization algorithms such as stochastic gradient descent (SGD). The optimization problems change slowly in the sense that the minimizers change at either a fixed or bounded rate. 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 needed from the distributions underlying each problem in order to ensure that the excess risk, i.e., 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. Experiments with synthetic and real data are used to confirm that this approach performs well.

Index Terms— stochastic optimization, gradient methods, machine learning, adaptive algorithms

1. INTRODUCTION

Consider solving a sequence of machine learning problems such as regression or classification by minimizing the expected value of a fixed loss function $\ell(\boldsymbol{x},\boldsymbol{z})$ at each time n:

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

For regression, $\boldsymbol{z}_n$ corresponds to the predictors and response at time n and $\boldsymbol{x}$ parameterizes the regression model. For classification, $\boldsymbol{z}_n$ corresponds to the features and label at time n and $\boldsymbol{x}$ parameterizes the classifier. Although, motivated by regression and classification, our framework works for any loss function $\ell(\boldsymbol{x},\boldsymbol{z})$ that satisfies certain properties discussed later. In the learning context, a *task* consists of the loss function $\ell(\boldsymbol{x},\boldsymbol{z})$ and the distribution p_n, and so our problem can be viewed as learning a sequence of tasks.

The problems change slowly at a constant but unknown rate in the sense that

$$\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\| = \rho \qquad \forall n \geq 2 \tag{2}$$

with $\boldsymbol{x}_n^*$ the minimizer of $f_n(\boldsymbol{x})$. In an extended version of this paper [1], we also consider slow changes at a bounded but unknown rate

$$\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\| \leq \rho \qquad \forall n \geq 2 \tag{3}$$

This work was supported by the NSF under award CCF 11-11342 through the University of Illinois at Urbana-Champaign.

Under this model, we sequentially find approximate minimizers $\boldsymbol{x}_n$ of each function $f_n(\boldsymbol{x})$ using K_n samples $\{\boldsymbol{z}_n(k)\}_{k=1}^{K_n} \overset{\text{iid}}{\sim} p_n$ from p_n by applying an optimization algorithm such as SGD starting from the previous approximate minimizer $\boldsymbol{x}_{n-1}$. We evaluate the quality of our approximate minimizers $\boldsymbol{x}_n$ through an excess risk criterion ε_n, i.e.,

$$\mathbb{E}\left[f_n(\boldsymbol{x}_n)\right] - f_n(\boldsymbol{x}_n^*) \leq \varepsilon_n$$

which is a standard criterion for optimization and learning problems [2]. Our goal is to determine adaptively the number of samples K_n required to achieve a desired excess risk ε for each n with ρ unknown. As ρ is unknown, we construct estimates of ρ, which, combined with properties of the chosen optimization algorithm, yield selection rules for the number of samples K_n required to achieve a target excess risk ε. Finally, we test our approach on synthetic and real data.

1.1. Related Work

Our problem has connections with *multi-task learning* (MTL) and *transfer learning*. In multi-task learning, one tries to learn several tasks simultaneously as in [3], [4], and [5] 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 [6], [7]. Multi-task learning could be applied to our problem by running a 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.

In *online optimization*, a sequence of functions $f_n(\boldsymbol{x})$ arrive, and the goal is to minimize the regret [8–17]. The idea of controlling the variation of the sequence of functions has been studied in [18] and [19]. In [19], the assumption on how the arriving functions change is equivalent to bounding

$$\sum_{n=2}^{T} \|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\|_2^2 \leq G_b.$$

Therefore, the work in [19] studies the regret while controlling the total variation in the optimal solutions over T time instants. In contrast, we control the variation of the optimal solutions at each time instant with (2) and then control the excess risk at each time instant.

In the *concept drift* problem, 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. 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 relevant model is *sequential supervised learning* (see [22]) in which we observe a stream of data consisting of feature/label pairs $(\boldsymbol{w}_n, y_n)$ at time n, with $\boldsymbol{w}_n$ being the feature vector and y_n being the label. At time n, we want to predict y_n given $\boldsymbol{w}_n$. Approaches based on sliding windows of L consecutive pairs [23, 24] and hidden Markov models (HMM) [25] have been studied.

None of the prior work discussed in this section involves 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.

2. ADAPTIVE SEQUENTIAL OPTIMIZATION WITH ρ KNOWN

For analysis, we assume that $\text{diam}(\mathscr{X}) < +\infty$ and the following assumptions on our functions $f_n(\boldsymbol{x})$ and the optimization algorithm:

A.1 For the optimization algorithm under consideration, there is a bound $b(d_0, K_n)$ such that

$$\mathbb{E}\left[f_n(\boldsymbol{x}_n)\right] - f_n(\boldsymbol{x}_n^*) \leq b(d_0, K_n)$$

with K_n the number of samples from p_n and $\mathbb{E}\|\boldsymbol{x}_n(0) - \boldsymbol{x}_n^*\|^2 \leq d_0$, where $\boldsymbol{x}_n(0)$ is the initial point of the optimization algorithm at time n. Finally, $b(d_0, K_n)$ is non-decreasing in d_0.

A.2 Each loss function $\ell(\boldsymbol{x}, \boldsymbol{z})$ is differentiable in $\boldsymbol{x}$. Each $f_n(\boldsymbol{x})$ is strongly convex with parameter m, i.e.,

$$f_n(\boldsymbol{y}) \geq f_n(\boldsymbol{x}) + \langle \nabla_{\boldsymbol{x}} f_n(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \tfrac{1}{2} m \|\boldsymbol{y} - \boldsymbol{x}\|^2$$

A.3 We can find initial points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ that satisfy the excess risk criterion with ε_1 and ε_2 known, i.e.,

$$\mathbb{E}\left[f_i(\boldsymbol{x}_i)\right] - f_i(\boldsymbol{x}_i^*) \leq \varepsilon_i \quad i = 1, 2$$

Remarks: For assumption A.1, we assume that the bound $b(d_0, K_n)$ depends on the number of samples K_n and not the number of iterations. For SGD, generally the number of iterations equals K_n as each sample is used to produce a noisy gradient. As an example, for SGD with a constant step size $\mu = 1/\sqrt{K}$, it holds that $b(d_0, K) = C_1 d_0/\sqrt{K} + C_2/\sqrt{K}$ for closed form constants C_1 and C_2 [26]. The extended paper [1] contains several collected examples of $b(d_0, K)$. In addition, we set $\boldsymbol{x}_n(0) = \boldsymbol{x}_{n-1}$ meaning that we use the approximate minimizer at time $n-1$ as the starting point for the new approximate minimizer at time n. For assumption A.3, we can fix K_i and set $\varepsilon_i = b(\text{diam}(\mathscr{X})^2, K_i)$ for $i = 1, 2$. Finally, under these assumptions, a slow changing assumption on $f_n(\boldsymbol{x})$ instead of $\boldsymbol{x}_n^*$, i.e., $f_n(\boldsymbol{x}_{n-1}^*) - f_n(\boldsymbol{x}_n^*) = \bar{\rho}$, can be converted to a bounded change condition on $\boldsymbol{x}_n^*$ as in (3) by exploiting strong convexity [27]:

$$\|\boldsymbol{x}_n^* - \boldsymbol{x}_{n-1}^*\| \leq \frac{2}{m}\left(f_n(\boldsymbol{x}_{n-1}^*) - f_n(\boldsymbol{x}_n^*)\right) = \frac{2\bar{\rho}}{m}$$

This shows that placing an assumption on the change in $\boldsymbol{x}_n^*$ is natural.

Now, we examine the case when the change in minimizers, ρ in (2) or (3), is known. The analysis in this section is the same under (2) or (3). We want to construct a bound ε_n on the excess risk at time n in terms of K_n and ρ, i.e., ε_n such that $\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq \varepsilon_n$. The idea is to start with the bounds from assumption A.3 and proceed inductively using the previous ε_{n-1} and ρ from (2). Suppose that ε_{n-1} bounds the excess risk at time $n-1$. Using the triangle inequality, strong convexity, Jensen's inequality, and (2) we have

$$\mathbb{E}\|\boldsymbol{x}_{n-1} - \boldsymbol{x}_n^*\|^2 \leq \left(\sqrt{\frac{2\varepsilon_{n-1}}{m}} + \rho\right)^2 \tag{4}$$

In comparison, we could use the estimate $\text{diam}^2(\mathscr{X})$ to bound $\mathbb{E}\|\boldsymbol{x}_{n-1} - \boldsymbol{x}_n^*\|^2$ and select K_n. If the bound in (4) is much smaller than $\text{diam}(\mathscr{X})^2$, then we need significantly fewer samples K_n to guarantee a desired excess risk. Now, by using the bound $b(d_0, K_n)$ from assumption A.1, we can set

$$\varepsilon_n \;=\; b\left(\left(\sqrt{\frac{2\varepsilon_{n-1}}{m}} + \rho\right)^2, K_n\right) \quad \forall n \geq 3$$

which yields 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 $\varepsilon_n \leq \varepsilon$ for all n, we set

$$K_1 = \min\{K \geq 1 \mid b\left(\text{diam}(\mathscr{X})^2, K\right) \leq \varepsilon\}$$

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

$$K^* = \min\left\{K \geq 1 \;\middle|\; b\left(\left(\sqrt{\frac{2\varepsilon}{m}} + \rho\right)^2, K\right) \leq \varepsilon\right\} \tag{5}$$

3. ESTIMATING ρ

In practice, we do not know ρ, and so we must construct an estimate $\hat{\rho}_n$ using the samples from each distribution p_n. We introduce one approach to estimate ρ under (2) and defer another approach and estimates under (3) to [1]. We show that for all n large enough with appropriately chosen sequences $\{t_n\}$, $\hat{\rho}_n + t_n \geq \rho$ almost surely. With this property, we will show that analysis similar to that in Section 2 holds.

3.1. Estimating One Step Change

First, we construct an estimate $\tilde{\rho}_i$ of the one step changes $\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\|$. Using the triangle inequality and variational inequalities from [28] yields

$$\begin{aligned}\|\boldsymbol{x}_i^* - \boldsymbol{x}_{i-1}^*\| &\leq \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\| + \|\boldsymbol{x}_i - \boldsymbol{x}_i^*\| + \|\boldsymbol{x}_{i-1} - \boldsymbol{x}_{i-1}^*\| \\ &\leq \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\| + \frac{1}{m}\|\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x}_i)\| + \frac{1}{m}\|\nabla_{\boldsymbol{x}} f_{i-1}(\boldsymbol{x}_{i-1})\|\end{aligned}$$

We then approximate $\|\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x}_i)\| = \|\mathbb{E}_{\boldsymbol{z}_i \sim p_i}\left[\nabla_{\boldsymbol{x}} \ell(\boldsymbol{x}_i, \boldsymbol{z}_i)\right]\|$ by

$$\hat{G}_i \triangleq \left\| \frac{1}{K_i} \sum_{k=1}^{K_i} \nabla_{\boldsymbol{x}} \ell(\boldsymbol{x}_i, \boldsymbol{z}_i(k)) \right\| \tag{6}$$

to yield the following estimate that we call the *direct estimate*:

$$\tilde{\rho}_i \triangleq \|\boldsymbol{x}_i - \boldsymbol{x}_{i-1}\| + \frac{1}{m}\|\hat{G}_i\| + \frac{1}{m}\|\hat{G}_{i-1}\| \tag{7}$$

3.2. Combining One Step Estimates

We average the one step estimates $\tilde{\rho}_i$ to yield a better estimate $\hat{\rho}_n = \frac{1}{n-1}\sum_{i=2}^{n}\tilde{\rho}_i$ of ρ at each time n under (2). The difficulty in analyzing the direct estimate comes because in approximating $\|\nabla f_i(\boldsymbol{x}_i)\|$ by $\|\hat{G}_i\|$ from (6), $\boldsymbol{x}_i$ is dependent on all the samples $\{\boldsymbol{z}_i(k)\}_{k=1}^{K_i}$, which rules out the use of simple concentration inequalities. For analysis, we need the following additional assumptions:

B.1 The loss function $\ell(\boldsymbol{x},\boldsymbol{z})$ has uniform Lipschitz continuous gradients in $\boldsymbol{x}$ with modulus L, i.e. ,

$$\|\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x},\boldsymbol{z}) - \nabla_{\boldsymbol{x}}\ell(\tilde{\boldsymbol{x}},\boldsymbol{z})\| \leq L\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|$$

B.2 Assuming $\mathscr{X}$ is d-dimensional, each component j of the gradient error $\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x},\boldsymbol{z}_n) - \nabla f_n(\boldsymbol{x})$ satisfies

$$\mathbb{E}\left[\exp\left\{s\left(\nabla_{\boldsymbol{x}}\ell(\boldsymbol{x},\boldsymbol{z}_n) - \nabla f_n(\boldsymbol{x})\right)_j\right\} \;\middle|\; \boldsymbol{x}\right] \leq \exp\left\{\tfrac{1}{2}\tfrac{C_g}{d^2}s^2\right\}$$

We show that $\hat{\rho}_n$ eventually upper bounds ρ. For analysis, we consider starting with $\boldsymbol{x}_{i-1}$ and producing $\tilde{\boldsymbol{x}}_i$ by the same process as the one that produced $\boldsymbol{x}_i$ except with an independent draw of samples $\{\tilde{\boldsymbol{z}}_i(k)\}_{k=1}^{K_i}$. Through this approach we obtain exponential concentration inequalities. Applying the Borel-Cantelli lemma then shows that eventually $\hat{\rho}_n$ plus a constant upper bounds ρ.

Theorem 1. *If B.1-B.2 hold and the sequence $\{t_n\}$ satisfies $\sum_{n=2}^{\infty} e^{-Cnt_n^2} < \infty$ for all $C > 0$, then for a sequence of constants $\{C_n\}$ and for all n large enough it holds that $\hat{\rho}_n + C_n + t_n \geq \rho$ almost surely.*

Proof. See [1]. □

3.3. Parameter Estimation

We may need to estimate parameters of the functions $\{f_n(\boldsymbol{x})\}$ such as the strong convexity parameter m to compute $b(d_0,K)$. Extensions to accomplish this are discussed in [1]. The analysis of parameter estimation is similar to the analysis of ρ estimation.

4. ADAPTIVE SEQUENTIAL OPTIMIZATION WITH ρ UNKNOWN

We now examine the case with ρ unknown. We extend the work of Section 2 using the estimates of ρ in Section 3. 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 $b(d_0,K_n) = \alpha(K_n)d_0 + \beta(K_n)$

We have demonstrated that assumption C.1 holds for the direct estimate of ρ (7) under (2). As long as C.1-C.2 hold, the analysis in this section is the same under (2) and (3). We first present a general result showing that for appropriate choices of K_n, the excess risk is well-behaved.

Theorem 2. *Under assumptions C.1- C.2 and with $K_n \geq K^*$ for all n large enough almost surely with K^* from* (5)*, we have* $\limsup_{n\to\infty}(\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*)) \leq \varepsilon$.

Proof. See [1]. □

4.1. Update Past Excess Risk Bounds

We first consider updating all past excess risk bounds as we go. At time n, we plug-in $\hat{\rho}_{n-1} + t_{n-1}$ in place of ρ and define for $i = 1,\ldots,n$

$$\hat{\varepsilon}_i^{(n)} \;=\; b\left(\left(\sqrt{\frac{2}{m}\hat{\varepsilon}_{i-1}^{(n)}} + (\hat{\rho}_{n-1} + t_{n-1})\right)^2, K_i\right)$$

If it holds that $\hat{\rho}_{n-1} + t_{n-1} \geq \rho$, then $\mathbb{E}[f_n(\boldsymbol{x}_n)] - f_n(\boldsymbol{x}_n^*) \leq \hat{\varepsilon}_n^{(i)}$ for $i = 1,\ldots,n$. Assumption C.1 guarantees that this holds for all n large enough almost surely. We can thus set K_n equal to the smallest K such that

$$b\left(\left(\sqrt{\frac{2}{m}\max\{\hat{\varepsilon}_{n-1}^{(n-1)}, \varepsilon\}} + (\hat{\rho}_{n-1} + t_{n-1})\right)^2, K\right) \leq \varepsilon$$

for all $n \geq 3$ to achieve excess risk ε. The maximum in this definition ensures that when $\hat{\rho}_{n-1} + t_{n-1} \geq \rho$, $K_n \geq K^*$ with K^* from (5). Therefore, we can apply Theorem 2.

4.2. Do Not Update Past Excess Risk Bounds

Updating all past estimates of the excess risk bounds from time 1 up to n imposes a computational and memory burden. We now analyze what happens when we do not update the past excess risk bounds. Suppose that for all $n \geq 3$ we set

$$K_n = \min\left\{K \geq 1 \;\middle|\; b\left(\left(\sqrt{\frac{2\varepsilon}{m}} + (\hat{\rho}_{n-1} + t_{n-1})\right)^2, K\right) \leq \varepsilon\right\} \tag{8}$$

This is the same form as the choice in (5) 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. Therefore, we can apply Theorem 2 again.

5. EXPERIMENTS

We focus on two regression applications here for synthetic and real data. Classification applications for synthetic and real data with support vector machines (SVM) are in [1].

5.1. Synthetic Regression

Consider a regression problem with synthetic data using the penalized quadratic loss $\ell(\boldsymbol{x},\boldsymbol{z}) = \frac{1}{2}\left(y - \boldsymbol{w}^\top\boldsymbol{x}\right)^2 + \frac{1}{2}\lambda\|\boldsymbol{x}\|^2$ with $\boldsymbol{z} = (\boldsymbol{w},y) \in \mathbb{R}^{d+1}$. The distribution of $\boldsymbol{z}_n$ is zero mean Gaussian with covariance matrix $\boldsymbol{\Sigma}_n$. Under these assumptions, we can analytically compute minimizers $\boldsymbol{x}_n^*$ of $f_n(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z}_n\sim p_n}[\ell(\boldsymbol{x},\boldsymbol{z}_n)]$. We change $\boldsymbol{\Sigma}_n$ appropriately to

ensure that $\|x_n^* - x_{n-1}^*\| = \rho$ holds for all n. We find approximate minimizers using SGD with $\lambda = 0.1$. We estimate ρ using the direct estimate. It can be checked that the assumptions are satisfied for both experiments considered in this section.

We let n range from 1 to 20 with $\rho = 1$, a target excess risk $\varepsilon = 0.1$, and K_n from (8). We average over twenty runs of our algorithm. Figure 1 shows $\hat{\rho}_n$, our estimate of ρ, which is above ρ in general. Figure 2 shows the number of samples K_n, which settles down. We can exactly compute $f_n(x_n) - f_n(x_n^*)$, and so by averaging over the twenty runs of our algorithm, we can estimate the excess risk (denoted "sample average estimate"). Figure 3 shows this estimate of the excess risk, the target excess risk, and our bound on the excess risk from Section 4.2. We achieve at least our targeted excess risk

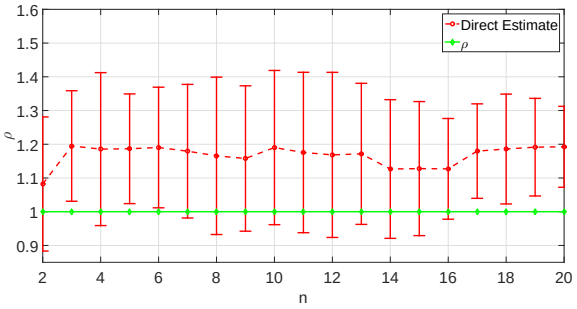

Fig. 1: ρ Estimate

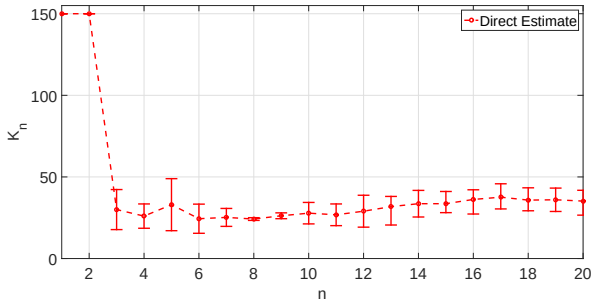

Fig. 2: K_n

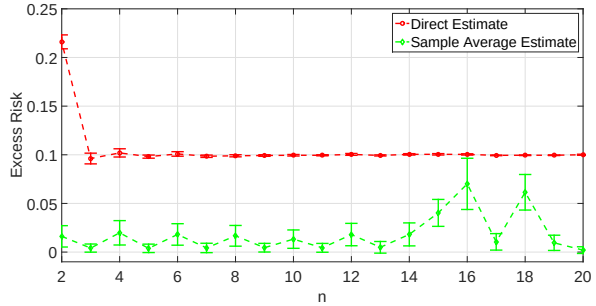

Fig. 3: Excess Risk

5.2. Panel Study on Income Dynamics Income - Regression

The Panel Study of Income Dynamics (PSID) surveyed individuals every year to gather demographic and income data annually from 1981-1997 [29]. We want to predict an individual's annual income (y) from several demographic features (w) including age, education, work experience, etc. chosen based on previous economic studies in [30]. Conceptually, the idea behind this experiment 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. The income is adjusted for inflation to 1997 dollars with mean \$20,294. We average over twenty runs of our algorithm by resampling without replacement [31]. 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 no other methods to select the number of samples K_n to control the excess risk against which we could compare our approach.

Figure 4 shows the test losses over time evaluated over twenty percent of the available samples. The test loss for our approach is substantially less than taking the same number of samples up front. The square root of the average test loss over this time period for our approach and all samples up front are $\$1153 \pm 352$ and $\$2805 \pm 424$ respectively in 1997 dollars.

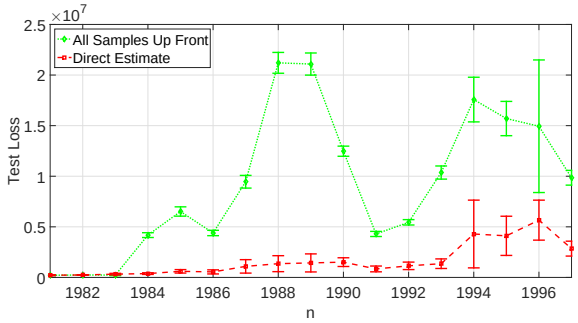

Fig. 4: Test Loss

6. CONCLUSION

We introduced a framework for adaptively solving a sequence of optimization problems with applications to machine learning. We developed estimates of the change in the minimizers used to determine the number of samples K_n needed to achieve a target excess risk ε. Experiments with synthetic and real data demonstrate that this approach is effective.

7. REFERENCES

[1] C. Wilson and V.V. Veeravalli, "Adaptive sequential optimization with applications to machine learning," *arXiv:1509.07422*, Sep. 2015.

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, The MIT Press, 2012.

[3] A. Agarwal, H. Daumé, and S. Gerber, "Learning multiple tasks using manifold regularization.," in *NIPS*, 2011, pp. 46–54.

[4] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2004, KDD '04, pp. 109–117, ACM.

[5] Y. Zhang and D. Yeung, "A convex formulation for learning task relationships in multi-task learning," *CoRR*, vol. abs/1203.3536, 2012.

[6] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct 2010.

[7] A. Agarwal, A. Rakhlin, and P. Bartlett, "Matrix regularization techniques for online multitask learning," Tech. Rep. UCB/EECS-2008-138, EECS Department, University of California, Berkeley, Oct 2008.

[8] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, Cambridge University Press, New York, N.Y., USA, 2006.

[9] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[10] J. Duchi and Y. Singer, "Efficient online and batch learning using forward backward splitting," *Journal of Machine Learning Research*, vol. 10, pp. 2899–2934, 2009.

[11] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Machine Learning*, vol. 69, pp. 169–192, 2007.

[12] E. Hazan P. Bartlett and A. Rakhlin, "Adaptive online gradient descent," in *Advances in Neural Information Processing Systems*, J.C. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20, pp. 65–72. MIT Press, Cambridge, MA, USA, 2008.

[13] S. Shalev-Shwartz and S.M. Kakade, "Mind the duality gap: Logarithmic regret algorithms for online optimization," in *Advances in Neural Information Processing Systems*, D. Koller, D.Schuurmans, Y. Bengio, and L. Bottou, Eds., MIT Press, 2009, vol. 21, pp. 1457–1464.

[14] S. Shalev-Shwartz and Y. Singer, "Convex repeated games and Fenchel duality," in *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, USA, 2006, vol. 19, pp. 1265–1271.

[15] S. Shalev-Shwartz and Y. Singer, "Logarithmic regret algorithms for strongly convex repeated games," In The Hebrew University, 2007.

[16] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," Tech. Rep., Microsoft, March 2010, no. MSR-TR-2010-23.

[17] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proceedings of the 20th International Conference on Machine Learning (ICML)*, 2003, pp. 928–936.

[18] A. Rakhlin and K. Sridharan, "Online Learning with Predictable Sequences," *ArXiv*, vol. abs/1208.3728, Aug. 2012.

[19] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu, "Online optimization with gradual variations," in *COLT*, 2012.

[20] Z. Towfic, J. Chu, and A. Sayed, "Online distirubted online classifcation in the midst of concept drifts," *Neurocomputing*, vol. 112, pp. 138–152, 2013.

[21] C. Tekin, L. Canzian, and M. van der Schaar, "Context adaptive big data stream mining," in *Allerton Conference*, 2014, pp. 46–54.

[22] T. Dietterich, "Machine learning for sequential data: A review," in *Structural, Syntactic, and Statistical Pattern Recognition*, 2002, pp. 15–30.

[23] T. Fawcett and F. Provost, "Adaptive fraud detection.," *Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 291–316, 1997.

[24] N. Qian and T. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of Molecular Biology*, vol. 202, pp. 865–884, Aug 1988.

[25] Y. Bengio and P. Frasconi, "Input-output HMM's for sequence processing," *IEEE Transactions on Neural Networks*, vol. 7(5), pp. 1231–1249, 1996.

[26] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, pp. 1574–1609, 2009.

[27] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.

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

[29] "Panel study of income dynamics: public use dataset," *Survey Research Center*, 2015.

[30] K. Murphy and F. Welch, "Empirical age-earnings profiles," *Journal of Labor Economics*, vol. 8, no. 2, pp. 202–29, 1990.

[31] T. Hastie, R. Tibshirani, and J.H. Friedman, *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*, New York: Springer-Verlag, 2001.