Stochastic Convex Optimization

Shai Shalev-Shwartz Nathan Srebro
Karthik Sridharan
Toyota Technological Institute–Chicago

June 2008

Abstract

Recently regret bounds for online convex optimization have been derived under very general conditions. These results can be used also in the stochastic batch setting by applying online-to-batch conversions. In this paper we study whether stochastic guarantees can be obtained more directly, for example using uniform convergence guarantees. We discover a surprising and complex situation: although the stochastic convex optimization problem is solvable (e.g. using online-to-batch conversions), no uniform convergence holds in the general case, and empirical minimization might fail. Rather then being a difference between online methods and a global minimization approach, we show that the key ingredient is strong convexity and regularization. Using stability arguments, we prove that strongly convex problems are solvable using empirical minimization. We then understand how weakly convex problems can be solved using regularization, and discuss how online algorithms can also be understood in terms of regularization.

1 Introduction

We consider the stochastic convex minimization problem

$$\operatorname*{argmin}_{\mathbf{w}\in\mathbf{W}} F(\mathbf{w}) \tag{1}$$

where $F(\mathbf{w}) = \mathbb{E}_\theta\left[f(\mathbf{w};\theta)\right]$ is the expectation of a random objective with respect to θ. The optimization is based on an i.i.d. sample $\theta_1,\ldots,\theta_n$ drawn from an unknown distribution. The goal is to choose $\mathbf{w}$ based on the sample and full knowledge of $f(\cdot,\cdot)$ and $\mathbf{W}$ so as to minimize $F(\mathbf{w})$. A special case is the familiar prediction setting where $\theta = (\mathbf{x}, y)$ is an instance-label pair and, e.g., $f(\mathbf{w};\mathbf{x},y) = \ell(\langle\mathbf{w},\phi(\mathbf{x})\rangle, y)$ for some convex loss function ℓ.

The situation in which the stochastic dependence on $\mathbf{w}$ is linear, as in the preceding example, is fairly well understood. When the domain $\mathbf{W}$ and the mapping ϕ are bounded, one can uniformly bound the deviation between the expected objective $F(\mathbf{w})$ and the empirical average

$$\hat{F}(\mathbf{w}) = \hat{\mathbb{E}}\left[f(\mathbf{w};\theta)\right] = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w};\theta_i). \tag{2}$$

This uniform convergence of $\hat{F}(\mathbf{w})$ to $F(\mathbf{w})$ justifies choosing the empirical minimizer

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \hat{F}(\mathbf{w}), \tag{3}$$

and guarantees that the expected value of $F(\hat{\mathbf{w}})$ converges to the optimal value $F(\mathbf{w}^\star)$, where $\mathbf{w}^\star = \arg\min_{\mathbf{w}} F(\mathbf{w})$ is the population optimum.

Our goal here is to consider the stochastic convex optimization problem more broadly, without assuming any metric or other structure on the parameter θ or mappings of it, or any special structure of the objective function $f(\cdot;\cdot)$.

An online analogue of this setting has recently received considerable attention. Online convex optimization concerns a sequence of convex functions $f(\cdot;\theta_1), \ldots, f(\cdot;\theta_n)$, which can be chosen by an adversary, and a sequence of online predictors $\mathbf{w}_i$, where $\mathbf{w}_i$ can depend only on $\theta_1, \ldots, \theta_{i-1}$. Online guarantees provide an upper bound on the online regret, $\frac{1}{n}\sum_i f(\mathbf{w}_i;\theta_i) - \min_{\mathbf{w}} \frac{1}{n}\sum_i f(\mathbf{w};\theta_i)$. Note the difference versus the stochastic setting, where we seek a *single predictor* $\tilde{\mathbf{w}}$ and would like to bound the *population sub-optimality* $F(\tilde{\mathbf{w}}) - F(\mathbf{w}^\star)$.

Zinkevich [1] showed that requiring $f(\mathbf{w};\theta)$ be Lipschitz-continuous w.r.t. $\mathbf{w}$ is enough for obtaining an online algorithm with online regret which diminishes as $\mathcal{O}\left(1/\sqrt{n}\right)$. If $f(\mathbf{w},\theta)$ is not merely convex w.r.t. $\mathbf{w}$, but also strongly convex, the regret bound can be improved to $\tilde{O}(1/n)$ [2].

These online results parallel known results in the stochastic setting, *when the stochastic dependence on* $\mathbf{w}$ *is linear*. However, they apply also in a much broader setting, when the dependence on $\mathbf{w}$ is not linear. E.g. when $f(\mathbf{w};\theta) = \|\mathbf{w}-\theta\|_p$ for $p \neq 2$. The requirement that the functions $\mathbf{w} \mapsto f(\mathbf{w};\theta)$ be Lipschitz-continuous is much more general than a specific requirement on the structure of the functions, and does not at all constrain the relationship between the functions. We note that this is quite different from the work of von Luxburg and Bousquet [3] who studied learning with functions that are Lipschitz with respect to θ.

The results for the online setting prompt us to ask whether similar results, requiring only Lipschitz continuity, can also be obtained for stochastic convex optimization. The answer we discover is surprisingly complex.

Our first surprising observation is that requiring Lipschitz continuity is *not* enough for ensuring uniform convergence of $\hat{F}(\mathbf{w})$ to $F(\mathbf{w})$, nor for the empirical minimizer $\hat{\mathbf{w}}$ to converge to an optimal solution. We present convex, bounded, Lipschitz-continuous examples where even as the sample size increases, the expected value of the empirical minimizer $\hat{\mathbf{w}}$ is bounded away from the population optimum: $F(\hat{\mathbf{w}}) = 1/2 > 0 = F(\mathbf{w}^\star)$.

In essentially all previously studied settings we are aware of where stochastic optimization is possible, we have at least some form of locally uniform convergence, and an empirical minimization approach is appropriate. In fact, for common models of supervised learning, it is known that uniform convergence is *equivalent* to stochastic optimization being possible [4]. This might lead us to think that Lipschitz-continuity is not enough to make stochastic convex optimization possible, even though it is enough to ensure online convex optimization is possible.

However, this gap between the online and stochastic setting cannot be, since it is possible to convert the online methods of Zinkevich and of Hazan *et al* to batch algorithms, with matching guarantees on the population sub-optimality $F(\tilde{\mathbf{w}}) - F(\mathbf{w}^\star)$. These guarantees hold for the specific output $\tilde{\mathbf{w}}$ of the algorithm, which is *not*, in general, the empirical minimizer. It seems, then, that we are in a strange situation where stochastic optimization is possible, but only using a specific (online) algorithm, rather than the more natural empirical minimizer.

We show that the "magic" can be understood not as a gap between online optimization and empirical minimization, but rather in terms of regularization. We first show that for a *strongly* convex stochastic optimization problem, even though we might still have no uniform convergence, the empirical minimizer is guaranteed to converge to the population optimum. This justifies stochastic convex optimization of general Lipschitz-continuous functions using *regularized* empirical minimiza-

tion. In fact, we discuss how Zinkevich's algorithm can also be understood in terms of minimizing an implicit regularized problem.

2 Setup and Background

A stochastic convex optimization problem is specified by a convex domain $\mathbf{W}$, which in this paper we always take to be a compact subset of a Hilbert space, and a function $f : \mathbf{W} \times \Theta \to \mathbb{R}$ which is convex w.r.t. its first argument. We say that the problem is "solvable" iff there exists a rule for choosing $\tilde{\mathbf{w}}$ based on an i.i.d. sample $\theta_1, \ldots, \theta_n$, and complete knowledge of $\mathbf{W}$ and $f(\cdot;\cdot)$, such that for any $\delta > 0$, any $\epsilon > 0$, and large enough sample size n, for any distribution over θ, with probability at least $1 - \delta$ over a sample of size n, we have $F(\tilde{\mathbf{w}}) - F(\mathbf{w}^\star) \leq \epsilon$.

We will consider various conditions on the convex optimization problem. We say that $\mathbf{W}$ is *bounded* by B if for all $\mathbf{w} \in \mathbf{W}$ we have $\|\mathbf{w}\| \leq B$. A function $f : \mathbf{W} \to \mathbb{R}$ is said to be L-*Lipschitz* if for any two vectors $\mathbf{w}_1, \mathbf{w}_2 \in \mathbf{W}$ we have $|f(\mathbf{w}_1) - f(\mathbf{w}_2)| \leq L \|\mathbf{w}_1 - \mathbf{w}_2\|$. We say that a function f is λ-*strongly convex* if for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathbf{W}$ and $\alpha \in [0, 1]$ we have

$$f(\alpha\mathbf{w}_1 + (1-\alpha)\mathbf{w}_2) \leq \alpha f(\mathbf{w}_1) + (1-\alpha) f(\mathbf{w}_2) - \frac{\lambda}{2}\alpha\,(1-\alpha)\,\|\mathbf{w}_1 - \mathbf{w}_2\|^2 \ . \tag{4}$$

Note that this strengthens the requirement that f is convex, which corresponds to setting $\lambda = 0$.

We say that a problem is a *generalized linear* problem if $f(\mathbf{w};\theta)$ can be written as

$$f(\mathbf{w}, \theta) = g(\langle \mathbf{w}, \phi(\theta) \rangle; \theta) + r(\mathbf{w}) \tag{5}$$

where $g : \mathbb{R} \times \Theta \to \mathbb{R}$ is convex w.r.t. its first argument, $r : \mathbf{W} \to \mathbb{R}$ is convex, and ϕ is an arbitrary mapping of θ to the Hilbert space in which $\mathbf{W}$ resides. A special case is supervised learning of a linear predictor with a convex loss function, where $g(\cdot;\cdot)$ encodes the loss function. Learnability results for linear predictors can in-fact be stated more generally as guarantees on stochastic optimization of generalized linear problems:

Theorem 1. *Consider a generalized linear stochastic convex optimization problem of the form* (5)*, such that the domain* $\mathbf{W}$ *is bounded by* B*, the image of* ϕ *is bounded by* R *and* $g(z;\theta)$ *is* L_g*-Lipschitz in* z*. Then for any distribution over* Θ *and any* $\delta > 0$*, with probability at least* $1 - \delta$*:*

$$\sup_{\mathbf{w}\in\mathbf{W}} \left| F(\mathbf{w}) - \hat{F}(\mathbf{w}) \right| \leq \mathcal{O}\left(\sqrt{\frac{B^2 (R L_g)^2 \log(1/\delta)}{n}} \right) \tag{6}$$

That is, the empirical values $\hat{F}(\mathbf{w})$ converge *uniformly*, for all $\mathbf{w} \in \mathbf{W}$, to their expectations $F(\mathbf{w})$. This ensures that with probability at least $1 - \delta$, for all $\mathbf{w} \in \mathbf{W}$:

$$F(\mathbf{w}) - F(\mathbf{w}^\star) \leq (\hat{F}(\mathbf{w}) - \hat{F}(\hat{\mathbf{w}})) + \mathcal{O}\left(\sqrt{\frac{B^2 (R L_g)^2 \log(1/\delta)}{n}} \right) \tag{7}$$

The empirical suboptimality term on the right-hand-side vanishes for the empirical minimizer $\hat{\mathbf{w}}$, establishing that empirical minimization solves the stochastic optimization problem with a rate of $\mathcal{O}\left(\sqrt{1/n}\right)$. Furthermore, (7) allows us to bound the population suboptimality in terms of the empirical suboptimality and obtain meaningful guarantees even for approximate empirical minimizers.

The non-stochastic term $r(\mathbf{w})$ does not play a role in the above bound, as it can always be canceled out. However, when this terms is strongly-convex (e.g. when it is a squared-norm regularization term, $r\mathbf{w} = \frac{\lambda}{2} \|w\|^2$), a faster convergence rate can be guaranteed:

Theorem 2. *[5] Consider a generalized linear stochastic convex optimization problem of the form* (5)*, such that $r(\mathbf{w})$ is λ-strongly convex and L_r-Lipschitz, the image of ϕ is bounded by R and $g(z;\theta)$ is L_g-Lipschitz in z. Then for any distribution over Θ and any $\delta > 0$, with probability at least $1-\delta$, for all $\mathbf{w} \in \mathbf{W}$:*

$$F(\mathbf{w}) - F(\mathbf{w}^\star) \leq 2(\hat{F}(\mathbf{w}) - \hat{F}(\hat{\mathbf{w}})) + \mathcal{O}\left(\frac{(RL_g + L_r)^2 \log(1/\delta)}{\lambda n}\right) \tag{8}$$

Online Convex Optimization

Zinkevich [1] established that Lipschitz continuity and convexity of the objective functions with respect to the optimization argument are sufficient for online optimization[1]:

Theorem 3. *[6, Corollary 1] Let $f : \mathbf{W} \times \Theta \to \mathbb{R}$ be such that $\mathbf{W}$ is bounded by B and $f(\mathbf{w}, \theta)$ is convex and L-Lipschitz with respect to $\mathbf{w}$. Then, there exists an online algorithm such that for any sequence $\theta_1, \ldots, \theta_n$ the sequence of online vectors $\mathbf{w}_1, \ldots, \mathbf{w}_n$ satisfies:*

$$\frac{1}{n}\sum_i f(\mathbf{w}_i;\theta_i) \leq \frac{1}{n}\sum_i f(\mathbf{w}^\star;\theta_i) + \mathcal{O}\left(\sqrt{\frac{B^2\,L^2}{n}}\right) \tag{9}$$

Subsequently, Hazan *et al* [2] showed that a faster rate can be obtained when the objective functions are not only convex, but also strongly convex:

Theorem 4. *[2, Theorem 1] Let $f : \mathbf{W} \times \Theta \to \mathbb{R}$ be such that function $f(\mathbf{w}, \theta)$ is λ-strongly convex and L-Lipschitz with respect to $\mathbf{w}$. Then, there exists an online algorithm such that for any sequence $\theta_1, \ldots, \theta_n$ the sequence of online vectors $\mathbf{w}_1, \ldots, \mathbf{w}_n$ satisfies:*

$$\frac{1}{n}\sum_i f(\mathbf{w}_i;\theta_i) \leq \frac{1}{n}\sum_i f(\mathbf{w}^\star;\theta_i) + \mathcal{O}\left(\frac{L^2\,\log(n)}{\lambda\, n}\right)$$

In this paper, we are not interested in the online setting, but rather in the batch stochastic optimization setting, where we would like to obtain a single predictor $\tilde{\mathbf{w}}$ with low expected value over *future* examples $F(\tilde{\mathbf{w}}) = \mathbb{E}_\theta\left[f(\tilde{\mathbf{w}};\theta)\right]$. Using martingale inequalities, it is possible to convert an online algorithm to a batch algorithm with a stochastic guarantee. One simple way to do so is to run the online algorithm on the stochastic sequence of functions $f(\cdot,\theta_1), \ldots, f(\cdot,\theta_n)$ and set the single predictor $\tilde{\mathbf{w}}$ to be the average of the online choices $\mathbf{w}_1, \ldots, \mathbf{w}_n$. Assuming the conditions of Theorem 3, it is possible to show (e.g. [7]) that with probability of at least $1-\delta$ we have

$$F(\tilde{\mathbf{w}}) - F(\mathbf{w}^\star) \;\leq\; \mathcal{O}\left(\sqrt{\frac{B^2\,L^2\,\log(1/\delta)}{n}}\right)\;. \tag{10}$$

[1] We present here slightly more general Theorem statements than those found in the original papers [1, 2]. We do not require differentiability, and instead of bounding the gradient and the Hessian we bound the Lipschitz constant and the parameter of strong convexity. The bound in Theorem 3 is also a bit tighter.

It is also possible to derive a similar guarantee assuming the conditions of Theorem 4:

$$F(\tilde{\mathbf{w}}) - F(\mathbf{w}^\star) \;\leq\; \mathcal{O}\left(\frac{L^2\,\log(n/\delta)}{\lambda\, n}\right). \tag{11}$$

The conditions for Theorem 3 generalize those of Theorem 1 when $r(\mathbf{w}) = 0$: If $f(\mathbf{w};\theta) = g(\langle \mathbf{w}, \phi(\theta)\rangle)$ satisfies the conditions of Theorem 1 then it also satisfies the conditions of Theorem 3 with $L = L_g R$ and the bound on the population sub-optimality of $\tilde{\mathbf{w}}$ given in (10) matches the guarantee on $\hat{\mathbf{w}}$ using Theorem 1. Similarly, the conditions of Theorem 4 generalize those of Theorem 2 with $L = RL_g + L_r$ and the guarantees are similar (except for a log-factor). It is important to note, however, that the guarantees (10) and (11) do *not* subsume Theorems 1 and 2, as the online-to-batch guarantees apply only to a specific choice $\tilde{\mathbf{w}}$ which is defined in terms of the behavior of a specific algorithms. They do not provide guarantees on the empirical minimizer, and certainly not a uniform guarantee in terms of the empirical sub-optimality.

3 Solvable, but not with Empirical Minimizer

The results of the previous section suggest that perhaps Lipschitz continuity is enough for obtaining guarantees on stochastic convex optimization using a more direct approach. In particular, that perhaps Lipschitz continuity is enough for ensuring uniform convergence, which in turn would imply that the empirical minimizer converges to the stochastic optimum, as in the linear case and in essentially all studied scenarios of stochastic optimization that we are aware of. Ensuring uniform convergence would further enables us to use approximate empirical minimizers, and bound the stochastic sub-optimality of *any* vector $\mathbf{w}$ in terms of its empirical sub-optimality, rather than obtaining a guarantee on the stochastic sub-optimality of only one specific procedural choice (obtained from running the online learning algorithm).

Unfortunately, this is not the case. Despite the fact that a bounded, Lipschitz-continuous, stochastic convex optimization problem is solvable, as demonstrated in the previous Section, we show here that uniform convergence does not hold and that it might not be solvable with empirical minimization.

Consider a convex stochastic optimization problem given by:

$$f_{(12)}(\mathbf{w};\theta) = \|\mathbf{w} * \theta\| \tag{12}$$

where for now we will set the domain to the d-dimensional unit sphere $\mathbf{W} = \left\{\mathbf{w} \in \mathbb{R}^d \,:\, \|\mathbf{w}\| \leq 1\right\}$ and take $\theta \in \Theta = \{0,1\}^d$ where $\mathbf{w} * \theta$ denotes an element-wise product. We will first consider a sequence of problems, where $d = 2^n$ for any sample size n, and later present the infinite-dimensional case. In any case the domain $\mathbf{W}$ is bounded by one, and for any θ the function $\mathbf{w} \mapsto f_{(12)}(\mathbf{w};\theta)$ is convex and 1-Lipschitz. Thus, the conditions of Theorem 3 hold, and the convex stochastic optimization problem is solvable by running Zinkevich's online algorithm and taking an average.

Consider a uniform distribution over Θ. For a random sample $\theta_1, \ldots, \theta_n$ we have that with probability greater than $1 - e^{-1} > 0.63$, there exists a coordinate $j \in 1 \ldots 2^n$ such that all parameter vectors θ_i in the sample are zero on the coordinate j, i.e. $\theta_i[j] = 0$. Let $\mathbf{e}_j \in \mathbf{W}$ be the standard basis vector corresponding to this coordinate. Then $\hat{F}_{(12)}(\mathbf{e}_j) = \frac{1}{n}\sum_i \|\mathbf{e}_j * \theta_i\| = \frac{1}{n}\sum_i |\theta_i[j]| = 0$ but $F_{(12)}(\mathbf{e}_j) = \mathbb{E}_\theta\left[\|\mathbf{e}_j * \theta\|\right] = \mathbb{E}_\theta\left[|\theta[j]|\right] = 1/2$. We established that for any n, we can construct a convex Lipschitz-continuous objective such that with probability at least 0.63 over the sample,

$\sup_{\mathbf{w}} \left| F_{(12)}(\mathbf{w}) - \hat{F}_{(12)}(\mathbf{w}) \right| \geq 1/2$. Furthermore, since $f(\cdot;\cdot)$ is non-negative, we have that $\mathbf{e}_j$ is an empirical minimizer, but its expected value $F_{(12)}(\mathbf{e}_j) = 1/2$ is far from the optimal expected value $\min_{\mathbf{w}} F_{(12)}(\mathbf{w}) = F_{(12)}(0) = 0$.

To formalize the example in a sample-size independent way, take $\mathbf{W}$ to be the unit sphere of a Hilbert space with orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots$, Θ be the set of infinite binary sequences, and $\mathbf{w} * \theta = \sum_j \theta[j] \langle \mathbf{w}, \mathbf{e}_j \rangle \mathbf{e}_j$. For any finite sample there is almost surely a coordinate j with $\theta_i[j] = 0$ for all i, and so we a.s. have an empirical minimizer $\hat{F}_{(12)}(\mathbf{e}_j) = 0$ with $F_{(12)}(\mathbf{e}_j) = 1/2 > 0 = F_{(12)}(0)$.

We see that although the stochastic convex optimization problem (12) is solvable (using Zinkevich's online algorithm), empirical minimization might not solve the problem!

It is also possible to construct a sharper counterexample, in which the *unique* empirical minimizer $\hat{\mathbf{w}}$ is far from having optimal expected value. To do so, we augment $f_{(12)}$ by a small term which ensures its empirical minimizer is unique:

$$f_{(13)}(\mathbf{w};\theta) = f_{(12)}(\mathbf{w};\theta) + 0.1 \left\| \mathbf{1} - \theta - \mathbf{w} \right\|_{\log_2 d} \tag{13}$$

where $\mathbf{1}$ is the vector of all ones and $\|\mathbf{x}\|_p$ denotes the ℓ_p-norm, and this time we will take $\mathbf{W} = \left\{ \mathbf{w} \in \mathbb{R}^d \middle| \|\mathbf{w}\| \leq 3 \right\}$. The problem is still convex, and 1.1-Lipschitz. The additional strictly convex term ensures the empirical minimizer is unique. Setting $d = 2^n$ as before ensures us that for a sample of size n, with probability greater than half, there are between one and eight "always zero" indices j with $\forall_i \theta_i[j] = 0$, and the second term ensures us that the unique empirical minimizer has $\hat{\mathbf{w}}[j] = 1$ on each one of these always zero coordinate. We then have $F_{(13)}(\hat{\mathbf{w}}) \geq F(\hat{\mathbf{w}}) \geq \mathbb{E}\left[|\theta[j]|\right] = \frac{1}{2}$ while $F_{(13)}(\mathbf{w}^\star) \leq F_{(13)}(0) \leq 0.1 \left\| \mathbf{1} \right\|_{\log_2 d} = 0.2$. And so, most of the time the empirical minimizer will not be close to the true optimum.

4 Empirical Minimization of a Strongly Convex Objective

We saw that empirical minimization is not adequate for stochastic convex optimization even if the objective is Lipschitz-continuous. We will now show that, if the objective $f(\mathbf{w};\theta)$ is *strongly* convex w.r.t. $\mathbf{w}$, the empirical minimizer *does* converge to the optimum. This is despite the fact that even in the strongly convex case, we still might not have uniform convergence of $\hat{F}(\mathbf{w})$ to $F(\mathbf{w})$.

Theorem 5. *Consider a stochastic convex optimization problem such that $f(\mathbf{w};\theta)$ is λ-strongly convex and L-Lipschitz with respect to $\mathbf{w}$. Let $\theta_1, \ldots, \theta_n$ be an i.i.d. sample and let $\hat{\mathbf{w}}$ be the empirical optimum. Then, with probability of at least $1 - \delta$ we have*

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star) \;\leq\; \mathcal{O}\left(\frac{L^2}{\delta\, \lambda\, n} \right). \tag{14}$$

Proof. The proof is based on the concept of uniform stability [8]. Denote

$$\hat{F}^{(i)}(\mathbf{w}) = \frac{1}{n-1} \sum_{j \neq i} f(\mathbf{w}, \theta_i)$$

the empirical average without the ith sample and let $\hat{\mathbf{w}}^{(i)} = \arg\min_{\mathbf{w}} = \hat{F}^{(i)}(\mathbf{w})$ be its minimizer. We first establish that the empirical minimizer is $\beta = \frac{2L^2}{\lambda n}$ uniformly stable, i.e. that

$\left|f(\hat{\mathbf{w}},\theta) - f(\hat{\mathbf{w}}^{(i)},\theta)\right| \leq \beta$ for all samples and all θ. To do so, we first calculate:

$$\begin{aligned}\hat{F}(\hat{\mathbf{w}}^{(i)}) - \hat{F}(\hat{\mathbf{w}}) &= \frac{f(\hat{\mathbf{w}}^{(i)},\theta_i) - f(\hat{\mathbf{w}},\theta_i)}{n} + \frac{\sum_{j\neq i}\left(f(\hat{\mathbf{w}}^{(i)},\theta_i) - f(\hat{\mathbf{w}},\theta_i)\right)}{n} \\ &= \frac{f(\hat{\mathbf{w}}^{(i)},\theta_i) - f(\hat{\mathbf{w}},\theta_i)}{n} + \frac{n-1}{n}\left(\hat{F}^{(i)}(\hat{\mathbf{w}}^{(i)}) - \hat{F}^{(i)}(\hat{\mathbf{w}})\right) \\ &\leq \frac{|f(\hat{\mathbf{w}}^{(i)},\theta_i) - f(\hat{\mathbf{w}},\theta_i)|}{n} \leq \frac{L}{n}\|\hat{\mathbf{w}}^{(i)} - \hat{\mathbf{w}}\|\,, \end{aligned} \tag{15}$$

where the first inequality follows from the fact that $\hat{\mathbf{w}}^{(i)}$ is the minimizer of $\hat{F}^{(i)}(\mathbf{w})$ and in the second inequality we use the Lipschitz property. But from strong convexity of $\hat{F}(\mathbf{w})$ and the fact that $\hat{\mathbf{w}}$ minimizes $\hat{F}(\mathbf{w})$ we also have that $\frac{\lambda}{2}\left\|\hat{\mathbf{w}}^{(i)} - \hat{\mathbf{w}}\right\|^2 \leq \hat{F}(\hat{\mathbf{w}}^{(i)}) - \hat{F}(\hat{\mathbf{w}})$. Combining this with (15) we obtain $\left\|\hat{\mathbf{w}}^{(i)} - \hat{\mathbf{w}}\right\| \leq 2L/(\lambda n)$ and from Lipschitz continuity we get

$$\left|f(\hat{\mathbf{w}},\theta) - f(\hat{\mathbf{w}}^{(i)},\theta)\right| \leq \tfrac{2L^2}{\lambda n} = \beta\,. \tag{16}$$

Now, from [8, Page 508] we have:

$$\mathbb{E}\left[F(\hat{\mathbf{w}}) - \hat{F}(\hat{\mathbf{w}})\right] \leq 2\beta = \tfrac{4L^2}{\lambda n} \tag{17}$$

Adding $\mathbb{E}\left[\hat{F}(\mathbf{w}^\star) - F(\mathbf{w}^\star)\right] = 0$ to the left-hand side and using the fact that $\hat{\mathbf{w}}$ minimizes $\hat{F}$:

$$\begin{aligned}\frac{4L^2}{\lambda n} &\geq \mathbb{E}\left[F(\hat{\mathbf{w}}) - \hat{F}(\hat{\mathbf{w}})\right] \\ &= \mathbb{E}\left[F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star)\right] + \mathbb{E}\left[\hat{F}(\mathbf{w}^\star) - \hat{F}(\hat{\mathbf{w}})\right] \geq \mathbb{E}\left[F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star)\right].\end{aligned}$$

Now, since the random variable $F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star)$ is non-negative we can apply Markov's inequality to get that

$$P\left[F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star) > \epsilon\right] \leq \frac{\mathbb{E}\left[F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star)\right]}{\epsilon} \leq \frac{4L^2}{\lambda\epsilon n}\,.$$

The proof follows by rearranging the above. □

We believe the dependence on δ in the above bound can be improved to $\log 1/\delta$, matching the online-to-batch guarantee (11).

We now turn to ask whether the convergence of the empirical minimizer in this case is a result of uniform convergence, and whether we can obtain a uniform bound in terms of the empirical sub-optimality as in (8). We first note that merely due to the fact that the empirical objective $\hat{F}$ is strongly convex, any approximate empirical minimizer must be close to $\hat{\mathbf{w}}$, and due to the fact that the expected objective F is Lipschitz-continuous any vector close to $\hat{\mathbf{w}}$ cannot have a much worse value than $\hat{\mathbf{w}}$. We therefore have, under the conditions of Theorem 5, that with probability at least $1-\delta$, for all $\mathbf{w} \in \mathbf{W}$:

$$F(\mathbf{w}) - F(\mathbf{w}^\star) \leq \sqrt{\frac{2L^2}{\lambda}}\sqrt{\hat{F}(\mathbf{w}) - \hat{F}(\hat{\mathbf{w}})} + \mathcal{O}\left(\frac{L^2}{\delta\,\lambda\,n}\right) \tag{18}$$

This is an immediate consequence of (14) and does not involve any stochastic properties of $\hat{F}$ and F. Although this uniform inequality does allow us to bound the population sub-optimality in terms of the empirical sub-optimality, the empirical sub-optimality must be quadratic in the desired population sub-optimality. Compare this dependence with the simple linear dependence of (8). Unfortunately, as we show next, this is the best that can be ensured.

To establish that the dependence on the empirical sub-optimality $\epsilon = \hat{F}(\mathbf{w}) - \hat{F}(\hat{\mathbf{w}})$ in (18) is tight, consider augmenting the objective function $f_{(12)}$ of Section 3 with a strongly convex term:

$$f_{(19)}(\mathbf{w};\theta) = f_{(12)}(\mathbf{w};\theta) + \frac{\lambda}{2}\left\|\mathbf{w}\right\|^2 \ . \tag{19}$$

The modified objective $f_{(19)}(\cdot;\cdot)$ is λ-strongly convex and $(1+\lambda)$-Lipschitz and thus satisfies the conditions of Theorem 5. For any sample, the unique empirical minimizer is the zero vector, which is also the population optimum. But consider a vector $t\mathbf{e}_j$ where j is some coordinate that is always zero on the sample (i.e. $\forall_i \theta_i[j] = 0$) and $t > 0$ is a scalar. We have that $\hat{F}_{(19)}(t\mathbf{e}_j) - \hat{F}_{(19)}(\hat{\mathbf{w}}) = \frac{\lambda}{2}t^2$ and so setting $t = \sqrt{2\epsilon/\lambda}$, we get an ϵ-empirical-suboptimal vector with population sub-optimality $F_{(19)}(t\mathbf{e}_j) - F_{(19)}(0) = \frac{1}{2}t + \frac{\lambda}{2}t^2 = \sqrt{\frac{\epsilon}{2\lambda}} + \epsilon$. This establishes that the dependence on $\sqrt{\frac{\epsilon}{\lambda}}$ in the first term of (18) is tight, and the situation is qualitatively different than the generalized linear case.

The above calculation also enables us establish that, even for a strongly convex objective, although the empirical minimizer itself does converge, not only might we not have uniform convergence of $\hat{F}(\mathbf{w})$ to $F(\mathbf{w})$, but we might not even have *local* uniform convergence. That is, we might not have $\sup_{\mathbf{w}} \left|\hat{F}(\mathbf{w}) - F(\mathbf{w})\right| \stackrel{n\to\infty}{\longrightarrow} 0$ even when the supremum is only over an arbitrarily small neighborhood of the optimum. This is in sharp contrast to essentially all other results on stochastic optimization that we are aware of.

5 Regularization

We now get back to the case where $f(\mathbf{w},\theta)$ is Lipschitz (and convex) w.r.t. $\mathbf{w}$ but not strongly convex. As we saw, empirical minimization may fail in this case, despite the guaranteed success of an online approach. Our goal in this section is to underscore a more direct, non-procedural, optimization criterion for stochastic optimization.

To do so, we define a regularized empirical minimization problem

$$\min_{\mathbf{w}\in\mathbf{W}} \left(\tfrac{\lambda}{2}\|\mathbf{w}\|^2 + \tfrac{1}{n}\sum_{i=1}^{n} f(\mathbf{w},\theta_i) \right) \ , \tag{20}$$

where λ is a parameter that will be determined later. The following theorem establishes that the minimizer of (20) is a good solution to the stochastic convex optimization problem:

Theorem 6. *Let $\mathbf{W}$ be a B bounded set and let $f : \mathbf{W}\times\Theta \to \mathbb{R}$ be such that for all $\theta\in\Theta$ the function $f(\mathbf{w},\theta)$ is convex and L-Lipschitz with respect to $\mathbf{w}$. Let $\theta_1,\ldots,\theta_n$ be an i.i.d. sample and let $\hat{\mathbf{w}}_\lambda$ be the minimizer of* (20) *with $\lambda = \sqrt{\frac{L^2}{\delta\, B^2\, n}}$. Then, with probability at least $1-\delta$ we have*

$$F(\hat{\mathbf{w}}_\lambda) - F(\mathbf{w}^\star) \ \leq \ \mathcal{O}\left(\sqrt{\frac{L^2\, B^2}{\delta\, n}}\right) .$$

Proof. Let $r(\mathbf{w},\theta) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + f(\mathbf{w},\theta)$ and let $R(\mathbf{w}) = \mathbb{E}_\theta\left[r(\mathbf{w},\theta)\right]$. From Theorem 5 we know that there exists a constant a such that

$$F(\hat{\mathbf{w}}) \;=\; \tfrac{\lambda}{2}\|\hat{\mathbf{w}}\|^2 + R(\hat{\mathbf{w}}) \;\;\leq\;\; \tfrac{\lambda}{2}\|\hat{\mathbf{w}}\|^2 + R(\mathbf{w}^\star) + \frac{a\,L^2}{\delta\,\lambda\,n} \;=\; \tfrac{\lambda}{2}\left(\|\hat{\mathbf{w}}\|^2 + \|\mathbf{w}^\star\|^2\right) + \frac{a\,L^2}{\delta\,\lambda\,n}$$

Using the boundedness assumption and plugging the value of λ we conclude our proof. □

From the above theorem and the discussion in Section 3 we conclude that regularization is a necessary tool for stochastic optimization. It is interesting to contrast this with the online learning algorithm of Zinkevich [1]. Seemingly, the online approach of Zinkevich does not rely on regularization. However, a more careful look reveals an underlying regularization also in the online technique. Indeed, Shalev-Shwartz [6] showed that Zinkevich's online learning algorithm can be viewed as approximate coordinate ascent optimization of the dual of the regularized problem (20). Furthermore, it is also possible to obtain the same online regret bound using a Follow-The-Regularized-Leader approach, which at each iteration i directly solves the regularized minimization problem (20) on $\theta_1,\ldots,\theta_{i-1}$. The key, then, seems to be regularization, rather then a procedural online versus global minimization approach.

Regularization vs Constraints

The role of regularization here is very different than in familiar settings such as ℓ_2 regularization in SVMs and ℓ_1 regularization in LASSO. In those settings regularization serves to constrain our domain to a low-complexity domain (e.g. low-norm predictors), where we rely on uniform convergence. In fact, almost all learning guarantees for such settings that we are aware of can be expressed in terms of some sort of uniform convergence. And as we mentioned, learnability (under the standard supervised learning model) is in fact *equivalent* to a uniform convergence property.

In our case, constraining the norm of $\mathbf{w}$ does *not* ensure uniform convergence. Consider the example $f_{(12)}(\cdot;\cdot)$ of Section 3. Even over a restricted domain $\mathbf{W}_r = \{\mathbf{w}|\|\mathbf{w}\| \leq r\}$, for arbitrarily small $r > 0$, the empirical averages $\hat{F}(\mathbf{w})$ do *not* uniformly converge to $F(\mathbf{w})$ and $\Pr\left(\lim_{n\to\infty}\sup_{\mathbf{w}\in\mathbf{W}_r}\left|\hat{F}(\mathbf{w}) - F(\mathbf{w})\right| > 0\right) = 1$. Furthermore, consider replacing the additional regularization term $\lambda\|\mathbf{w}\|^2$ with a constraint on the norm of $\|\mathbf{w}\|$, namely, solving the problem $\tilde{\mathbf{w}} = \arg\min_{\|\mathbf{w}\|\leq r}\hat{F}(\mathbf{w})$. As we show below, we cannot set r in a distribution-independent way (i.e. without knowing the solution...), as we did for λ. To see this, note that for the example $f_{(12)}(\cdot;\cdot)$ we must have $r \to 0$ to ensure $F(\tilde{\mathbf{w}}) \to F(\mathbf{w}^\star)$. However, for $f(\mathbf{w}) = \|\mathbf{e}_1 - \mathbf{w}\|$, we must set $r \to 1$. If the stochastic convex optimization problem includes both types of functions, no constraint will work for all distributions over functions. This sharply contrasts with traditional uses of regularization, were learning guarantees are actually typically stated in terms of a constraint on the norm rather than in terms of a regularization parameter.

6 Summary

Following the work of Zinkevich [1], we expected to be able to generalize well established results on stochastic optimization of linear functions also to the more general Lipschitz-convex case. We discovered a complex and unexpected situation, where strong convexity and regularization play a

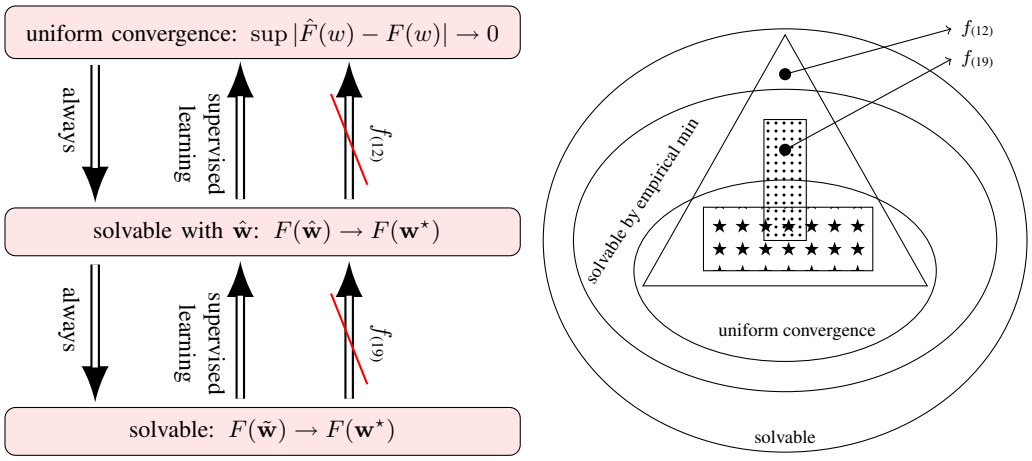

Figure 1: Left: Relationship between different properties of stochastic optimization problems. Right: Lipschitz-continuous convex problems (triangle) are all solvable, but not necessarily using empirical minimization. Lipschitz-continuous strongly convex problems (dotted rectangle) are all solvable with empirical minimization, but uniform convergence might not hold. For bounded generalized linear problems (starred rectangle), uniform convergence always holds. Our two separating examples are also indicated.

key role and ultimately did reach an understanding of stochastic convex optimization that does not rely on online techniques. Figure 1 summarizes some of our results.

For stochastic objectives that arise from supervised prediction problems, it is well known that learnability, i.e. solvability of the stochastic optimization problem, is equivalent to uniform convergence, and so whenever the problem is solvable, it is solvable using empirical minimization [4]. However, we demonstrated stochastic optimization problems in which these equivalences do not hold. There is no contradiction, since stochastic optimization problems that arise from supervised learning have a restricted structure, and in particular the examples we study are not among such problems. In fact, for a reasonable loss function, in order to make $f(\mathbf{w};\mathbf{x},y) = \ell(\text{pred}(\mathbf{w},\mathbf{x}),y)$ convex for both positive and negative labels, we must essentially make the prediction function $\text{pred}(\mathbf{w},\mathbf{x})$ both convex and concave in $\mathbf{w}$, i.e. linear. And so the only stochastic (or online) convex optimization problems that correspond to supervised problems are generalized linear problems.

References

[1] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[2] E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2006.

[3] Ulrike von Luxburg and Olivier Bousquet. Distance–based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5:669–695, 2004.

[4] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997.

[5] K. Sridharan. Fast convergence rates for excess regularized risk with application to SVM. `http://ttic.uchicago.edu/~karthik/con.pdf`, 2008.

[6] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

[7] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, September 2004.

[8] Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.