Stochastic Convex Optimization

Shai Shalev-Shwartz
TTI-Chicago
shai@tti-c.org

Ohad Shamir
The Hebrew University
ohadsh@cs.huji.ac.il

Nathan Srebro
TTI-Chicago
nati@uchicago.edu

Karthik Sridharan
TTI-Chicago
karthik@tti-c.org

Abstract

For supervised classification problems, it is well known that learnability is equivalent to uniform convergence of the empirical risks and thus to learnability by empirical minimization. Inspired by recent regret bounds for online convex optimization, we study stochastic convex optimization, and uncover a surprisingly different situation in the more general setting: although the stochastic convex optimization problem is learnable (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.

Our results demonstrate that the celebrated theorem of Alon *et al* on the equivalence of learnability and uniform convergence does not extend to Vapnik's General Setting of Learning, that in the General Setting considering only empirical minimization is not enough, and that despite Vanpnik's result on the equivalence of *strict* consistency and uniform convergence, uniform convergence is only a sufficient, but not necessary, condition for meaningful non-trivial learnability.

1 Introduction

We consider the stochastic convex minimization problem

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

where $F(\mathbf{w}) = \mathbb{E}_Z\left[f(\mathbf{w}; Z)\right]$ is the expectation, with respect to Z, of a random objective that is convex in $\mathbf{w}$. The optimization is based on an i.i.d. sample $z_1, \ldots, z_n$ drawn from an unknown distribution. The goal is to choose $\mathbf{w}$ based on the sample and full knowledge of $f(\cdot,\cdot)$ and $\mathcal{W}$ so as to minimize $F(\mathbf{w})$. Alternatively, we can also think of an unknown distribution over convex functions, where we are given a sample of functions $\{\mathbf{w} \mapsto f(\mathbf{w}; z_i)\}$ and would like to optimize the expected function. A special case is the familiar prediction setting where $z = (\mathbf{x}, y)$ is an instance-label pair, $\mathcal{W}$ is a subset of a Hilbert space, and $f(\mathbf{w}; \mathbf{x}, y) = \ell(\langle \mathbf{w}, \phi(\mathbf{x})\rangle, y)$ for some convex loss function ℓ and feature mapping ϕ.

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

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

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 $F(\mathbf{w}^\star) = \inf_{\mathbf{w}} F(\mathbf{w})$. Furthermore, a similar guarantee can also be obtained for an approximate minimizer of the empirical objective.

Our goal here is to consider the stochastic convex optimization problem more broadly, without assuming any metric or other structure on the parameter z or mappings of it, or any special structure of the objective function $f(\cdot;\cdot)$. Viewed as optimization based on a sample of functions, we do not impose any constraints on the functions, or the relationship between the functions, except that each function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ separately is convex and Lipschitz-continuous.

An online analogue of this setting has recently received considerable attention. Online convex optimization concerns a sequence of convex functions $f(\cdot; z_1), \ldots, f(\cdot; z_n)$, which can be chosen by an adversary, and a sequence of online predictors $\mathbf{w}_i$, where $\mathbf{w}_i$ can depend only on $z_1, \ldots, z_{i-1}$. Online guarantees provide an upper bound on the online regret, $\frac{1}{n}\sum_i f(\mathbf{w}_i; z_i) - \min_{\mathbf{w}} \frac{1}{n}\sum_i f(\mathbf{w}; z_i)$. 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 [Zin03] showed that requiring $f(\mathbf{w}; z)$ to be Lipschitz-continuous w.r.t. $\mathbf{w}$ is enough for obtaining an online algorithm with online regret which diminishes as $1/\sqrt{n}$. If $f(\mathbf{w}, z)$ is not merely convex w.r.t. $\mathbf{w}$, but also strongly convex, the regret diminishes with a faster rate of $1/n$ [HKKA06].

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 stochastic dependence on $\mathbf{w}$ is not linear, e.g. when $f(\mathbf{w};z) = \|\mathbf{w}-z\|_p$ for $p \neq 2$. The requirement that the functions $\mathbf{w} \mapsto f(\mathbf{w};z)$ 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. That is, we can think of z as parameterizing all possible Lipschitz-continuous convex functions $\mathbf{w} \mapsto f(\mathbf{w};z)$. We note that this is quite different from the work of von Luxburg and Bousquet [vLB04] who studied learning with functions that are Lipschitz with respect to z.

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 learning or 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 [ABCH97]. 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 method of Zinkevich to a batch algorithm, with a matching guarantee on the population sub-optimality $F(\tilde{\mathbf{w}}) - F(\mathbf{w}^\star)$. This guarantee holds 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.

To do so, 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 result seems to defy Vapnik's celebrated result on the equivalence of uniform convergence and *strict* consistency of the empirical minimizer [Vap95, Vap98]. We explain why there is no contradiction here: Vapnik's notion of "strict consistency" is too strict and does not capture all situations in which learning is non-trivial, yet still possible.

Convergence of the empirical minimizer to the population optimum for strongly convex objectives justifies stochastic convex optimization of weakly convex Lipschitz-continuous functions using *regularized* empirical minimization. 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 $\mathcal{W}$, which in this paper we always take to be a closed and bounded subset of a Hilbert space $\mathcal{H}$, and a function $f : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ which is convex w.r.t. its first argument.

- We say that the problem is **learnable** (or "solvable") iff there exists a rule for choosing $\tilde{\mathbf{w}}$ based on an i.i.d. sample $z_1, \ldots, z_n$, and complete knowledge of $\mathcal{W}$ and $f(\cdot;\cdot)$, such that for any $\delta > 0$, any $\epsilon > 0$, and large enough sample size n, for any distribution over z, with probability at least $1-\delta$ over a sample of size n, we have $F(\tilde{\mathbf{w}}) \leq F(\mathbf{w}^\star) + \epsilon$. We say that such a rule is **uniformly consistent**, or that it "solves" the stochastic optimization problem.
- We say that the problem is **bounded** by B iff for all $\mathbf{w} \in \mathcal{W}$ we have $\|\mathbf{w}\| \leq B$.
- We say that the problem is L**-Lipschitz** if $f(\mathbf{w};z)$ is L-Lipschitz w.r.t. $\mathbf{w}$. That is, for any $z \in \mathcal{Z}$ and $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ we have
$$|f(\mathbf{w}_1;z) - f(\mathbf{w}_2;z)| \leq L \left\|\mathbf{w}_1 - \mathbf{w}_2\right\| .$$
- We say that the problem λ**-strongly convex** if for any $z \in \mathcal{Z}$, $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ and $\alpha \in [0,1]$ we have
$$f(\alpha\mathbf{w}_1 + (1-\alpha)\mathbf{w}_2;z) \leq \alpha f(\mathbf{w}_1;z) + (1-\alpha) f(\mathbf{w}_2;z) - \tfrac{\lambda}{2}\alpha(1-\alpha)\left\|\mathbf{w}_1 - \mathbf{w}_2\right\|^2 .$$
Note that this strengthens the convexity requirement, which corresponds to setting $\lambda = 0$.

2.1 Generalized Linear Stochastic Optimization

We say that a problem is a *generalized linear* problem if $f(\mathbf{w};z)$ can be written as
$$f(\mathbf{w};z) = g(\langle \mathbf{w}, \phi(z) \rangle; z) + r(\mathbf{w}) \tag{4}$$
where $g : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$ is convex w.r.t. its first argument, $r : \mathcal{W} \to \mathbb{R}$ is convex, and $\phi : \mathcal{Z} \to \mathcal{H}$. 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* (4)*, such that the domain $\mathcal{W}$ is bounded by B, the image of ϕ is bounded by R and $g(u;z)$ is L_g-Lipschitz in u. Then for any distribution over z and any $\delta > 0$, with probability at least $1-\delta$ over a sample of size n:*
$$\sup_{\mathbf{w} \in \mathcal{W}} \left| F(\mathbf{w}) - \hat{F}(\mathbf{w}) \right| \leq \mathcal{O}\left(\sqrt{\frac{B^2 (R L_g)^2 \log(1/\delta)}{n}} \right)$$

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

$$F(\mathbf{w}) - F(\mathbf{w}^\star) \leq (\hat{F}(\mathbf{w}) - \hat{F}(\hat{\mathbf{w}})) + \mathcal{O}\left(\sqrt{\frac{B^2(RL_g)^2\log(1/\delta)}{n}}\right) \quad (5)$$

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 $\sqrt{1/n}$. Furthermore, (5) 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, $r(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2$), a faster convergence rate can be guaranteed:

Theorem 2. *[SSS08] Consider a generalized linear stochastic convex optimization problem of the form* (4)*, such that $r(\mathbf{w})$ is λ-strongly convex, the image of ϕ is bounded by R and $g(u;z)$ is L_g-Lipschitz in u. Then for any distribution over z and any $\delta > 0$, with probability at least $1-\delta$ over a sample of size n, for all $\mathbf{w} \in \mathcal{W}$:*

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

2.2 Online Convex Optimization

Zinkevich [Zin03] established that Lipschitz continuity and convexity of the objective functions with respect to the optimization argument are sufficient for online optimization[1]:

Theorem 3. *[Sha07, Corollary 1] Let $f : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ be such that $\mathcal{W}$ is bounded by B and $f(\mathbf{w}, z)$ is convex and L-Lipschitz with respect to $\mathbf{w}$. Then, there exists an online algorithm such that for any sequence $z_1, \ldots, z_n$ the sequence of online vectors $\mathbf{w}_1, \ldots, \mathbf{w}_n$ satisfies:*

$$\frac{1}{n}\sum_i f(\mathbf{w}_i; z_i) \leq \frac{1}{n}\sum_i f(\mathbf{w}^\star; z_i) + \mathcal{O}\left(\sqrt{\frac{B^2 L^2}{n}}\right) \quad (6)$$

Subsequently, Hazan *et al* [HKKA06] showed that a faster rate can be obtained when the objective functions are not only convex, but also strongly convex:

Theorem 4. *[HKKA06, Theorem 1] Let $f : \mathcal{W} \times \mathbb{Z} \to \mathbb{R}$ be such that function $f(\mathbf{w}, z)$ is λ-strongly convex and L-Lipschitz with respect to $\mathbf{w}$. Then, there exists an online algorithm such that for any sequence $z_1, \ldots, z_n$ the sequence of online vectors $\mathbf{w}_1, \ldots, \mathbf{w}_n$ satisfies:*

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

[1] We present here slightly more general theorem statements than those found in the original papers [Zin03, HKKA06]. 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 than that originally established by Zinkevich.

Online-to-batch conversions

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 $F(\tilde{\mathbf{w}}) = \mathbb{E}_z[f(\tilde{\mathbf{w}}; z)]$. 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 $f(\cdot, z_1), \ldots, f(\cdot, z_n)$ 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. [CCG04]) that with probability of at least $1-\delta$ we have

$$F(\tilde{\mathbf{w}}) - F(\mathbf{w}^\star) \;\leq\; \mathcal{O}\left(\sqrt{\frac{B^2\,L^2\,\log(1/\delta)}{n}}\right). \quad (7)$$

It is also possible to derive a similar guarantee assuming the conditions of Theorem 4 [KT08]:

$$F(\tilde{\mathbf{w}}) - F(\mathbf{w}^\star) \;\leq\; \mathcal{O}\left(\frac{L^2\,\log(n/\delta)}{\lambda\, n}\right). \quad (8)$$

The conditions for Theorem 3 generalize those of Theorem 1 when $r(\mathbf{w}) = 0$: If $f(\mathbf{w}; z) = g(\langle \mathbf{w}, \phi(z)\rangle)$ 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 (7) matches the guarantee on $\hat{\mathbf{w}}$ using Theorem 1. Similarly, the conditions of Theorem 4 roughly generalize those of Theorem 2 with $L = RL_g + L_r$ and the guarantees are similar (except for a log-factor, and as long as $L_r = \mathcal{O}(RL_g)$). It is important to note, however, that the guarantees (7) and (8) 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 algorithm. They do not provide guarantees on the empirical minimizer, and certainly not a uniform guarantee in terms of the empirical sub-optimality.

3 Warm-Up: Finite Dimensional Case

We begin by noting that in the finite dimensional case, Lipschitz continuity *is* enough to guarantee uniform convergence, hence also learnability via empirical minimization.

Theorem 5. *Let $\mathcal{W} \subset \mathbb{R}^d$ be bounded by B and let $f(\mathbf{w}, z)$ be L-Lipschitz w.r.t. $\mathbf{w}$. Then with probability of at least $1-\delta$ over a sample of size n, for all $\mathbf{w} \in \mathcal{W}$:*

$$\left|F(\mathbf{w}) - \hat{F}(\mathbf{w})\right| \leq \mathcal{O}\left(\sqrt{\frac{L^2 B^2 d\log(n)\log(\frac{d}{\delta})}{n}}\right)$$

Proof. We will show uniform convergence by bounding the ℓ_∞-covering number of the class of functions $\mathcal{F} = \{z \mapsto f(\mathbf{w}; z) | \mathbf{w} \in \mathcal{W}\}$. To do so, we first note that as a subset of an ℓ_2-sphere, we can bound the covering number of $\mathcal{W}$ with respect to the Euclidean distance $d_2(\mathbf{w}_1, \mathbf{w}_2) = \|\mathbf{w}_1 - \mathbf{w}_2\|$ [VG05]: (for $d > 3$)

$$\mathcal{N}(\epsilon, \mathcal{W}, d_2) = \mathcal{O}\left(d^2\left(\tfrac{B}{\epsilon}\right)^d\right) \quad (9)$$

We now turn to covering numbers of $\mathcal{F}$ with respect to the ℓ_∞ distance $d_\infty(f(\mathbf{w}_1;\cdot), f(\mathbf{w}_2;\cdot)) = \sup_z |f(\mathbf{w}_1;z) - f(\mathbf{w}_2;z)|$. By Lipschitz continuity, for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ we have $\sup_z |f(\mathbf{w}_1;z) - f(\mathbf{w}_2;z)| \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|$. An ϵ-covering of $\mathcal{W}$ w.r.t. d_2 therefore yields an $L\epsilon$-covering of $\mathcal{F}$ w.r.t. d_∞ distances, and so:

$$\mathcal{N}(\epsilon, \mathcal{F}, d_\infty) \leq \mathcal{N}(\epsilon/L, \mathcal{W}, d_2) = \mathcal{O}\Big(d^2 \left(\tfrac{LB}{\epsilon}\right)^d\Big) \quad (10)$$

Noting that the empirical ℓ_1 covering number is bounded by the d_∞ covering number, and using a uniform bound in terms of empirical ℓ_1 covering numbers we get [Pol84]:

$$\begin{aligned}\Pr\big(\sup_{\mathbf{w}\in\mathcal{W}} \Big|F(\mathbf{w}) - \hat{F}(\mathbf{w})\Big| \geq \epsilon\big) \\ \leq 8N(\epsilon, \mathcal{F}, d_\infty)\exp(-\tfrac{n\epsilon^2}{128LR}) \\ \leq \mathcal{O}\Big(d^2\left(\frac{LB}{\epsilon}\right)^d \exp(-\tfrac{n\epsilon^2}{128LR})\Big).\end{aligned}$$

Equating the right-hand-side to δ and bounding ϵ we get the bound in the Theorem. □

We can therefore conclude that empirical minimization is uniformly consistent with the same rate as in Theorem 5:

$$F(\hat{\mathbf{w}}) \leq F(\mathbf{w}^\star) + \mathcal{O}\left(\sqrt{\frac{L^2B^2d\log(n)\log(\frac{d}{\delta})}{n}}\right) \quad (11)$$

with probability at least $1-\delta$ over a sample of size n. This is the standard approach for establishing learnability. We now turn to ask whether such an approach can also be taken in the infinite dimensional case, i.e. yielding a bound that does not depend on the dimensionality.

4 Learnable, but not with Empirical Minimizer

The results of the Section 2.2 suggest that perhaps Lipschitz continuity is enough for obtaining guarantees on stochastic convex optimization using a more direct approach, even in infinite dimensions. In particular, that perhaps Lipschitz continuity is enough for ensuring uniform convergence, which in turn would imply learnability using empirical minimization, as in the infinite dimensional linear case, the finite dimensional Lipschitz case, and in essentially all studied scenarios of stochastic optimization that we are aware of. Ensuring uniform convergence would further enable 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 learnable even in infinite dimensions, as discussed in Section 2.2, we show here that uniform convergence does not hold and that it might not be learnable with empirical minimization.

4.1 Empirical Minimizer far from Population Optimal

Consider a convex stochastic optimization problem given by:

$$\begin{aligned}f_{(12)}(\mathbf{w};(\mathbf{x},\boldsymbol{\alpha})) &= \|\boldsymbol{\alpha} * (\mathbf{w} - \mathbf{x})\| \\ &= \sqrt{\sum_i \boldsymbol{\alpha}^2[i](\mathbf{w}[i] - \mathbf{x}[i])^2} \end{aligned} \quad (12)$$

where for now we will set the domain to the d-dimensional unit sphere $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq 1\}$ and take $z = (\mathbf{x}, \boldsymbol{\alpha})$ with $\boldsymbol{\alpha} \in [0,1]^d$ and $\mathbf{x} \in \mathcal{W}$, and where $u * v$ denotes an element-wise product. We will first consider a sequence of problems, where $d = 2^n$ for any sample size n, and establish that we cannot expect a convergence rate which is independent of the dimensionality d. We then formalize this example in infinite dimensions.

One can think of the problem (12) as that of finding the "center" of an unknown distribution over $\mathbf{x} \in \mathbb{R}^d$, where we also have stochastic per-coordinate "confidence" measures $\boldsymbol{\alpha}[i]$. We will actually focus on the case where some coordinates are missing, i.e. occasionally $\boldsymbol{\alpha}[i] = 0$.

In any case the domain $\mathcal{W}$ is bounded by one, and for any $z = (\mathbf{x}, \boldsymbol{\alpha})$ the function $\mathbf{w} \mapsto f_{(12)}(\mathbf{w}; z)$ is convex and 1-Lipschitz. Thus, the conditions of Theorem 3 hold, and the convex stochastic optimization problem is learnable by running Zinkevich's online algorithm and taking an average.

Consider the following distribution over $Z = (\mathbf{X}, \boldsymbol{\alpha})$: $\mathbf{X} = 0$ with probability one, and $\boldsymbol{\alpha}$ is uniform over $\{0,1\}^d$. That is, $\boldsymbol{\alpha}[i]$ are i.i.d. uniform Bernoulli. For a random sample $(\mathbf{x}_1, \boldsymbol{\alpha}_1), \ldots, (\mathbf{x}_n, \boldsymbol{\alpha}_n)$ 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 confidence vectors $\boldsymbol{\alpha}_i$ in the sample are zero on the coordinate j, i.e. $\boldsymbol{\alpha}_i[j] = 0$ for all $i = 1..n$. Let $\mathbf{e}_j \in \mathcal{W}$ be the standard basis vector corresponding to this coordinate. Then

$$\hat{F}_{(12)}(\mathbf{e}_j) = \frac{1}{n}\sum_i \|\boldsymbol{\alpha}_i * (\mathbf{e}_j - 0)\| = \frac{1}{n}\sum_i |\boldsymbol{\alpha}_i[j]| = 0$$

but

$$F_{(12)}(\mathbf{e}_j) = \mathbb{E}_{\mathbf{X},\boldsymbol{\alpha}}\left[\|\boldsymbol{\alpha} * (\mathbf{e}_j - 0)\|\right] = \mathbb{E}_{\mathbf{X},\boldsymbol{\alpha}}\left[|\boldsymbol{\alpha}[j]|\right] = 1/2.$$

We established that for any n, we can construct a convex Lipschitz-continuous objective in high enough dimension such that with probability at least 0.63 over the sample, $\sup_{\mathbf{w}} \Big|F_{(12)}(\mathbf{w}) - \hat{F}_{(12)}(\mathbf{w})\Big| \geq 1/2$. 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 $\min_{\mathbf{w}} F_{(12)}(\mathbf{w}) = F_{(12)}(0) = 0$.

4.2 In Infinite Dimensions: Population Minimizer Does Not Converge to Population Optimum

To formalize the example in a sample-size independent way, take $\mathcal{W}$ to be the unit sphere of an infinite-dimensional Hilbert space with orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots$, where for $\mathbf{v} \in \mathcal{W}$, we refer to its coordinates $\mathbf{v}[j] = \langle \mathbf{v}, \mathbf{e}_j \rangle$ w.r.t this basis. The confidences $\boldsymbol{\alpha}$ are now a mapping of each coordinate to $[0,1]$. That is, an infinite sequence of reals in $[0,1]$. The element-wise product operation $\boldsymbol{\alpha} * \mathbf{v}$ is defined with respect to this basis and the objective function $f_{(12)}$ of equation (12) is well defined in this infinite-dimensional space.

We again take a distribution over $Z = (\mathbf{X}, \boldsymbol{\alpha})$ where $\mathbf{X} = 0$ and $\boldsymbol{\alpha}$ is an i.i.d. sequence of uniform Bernoulli random variables. Now, for any finite sample there is almost surely a coordinate j with $\boldsymbol{\alpha}_i[j] = 0$ 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 learnable (using Zinkevich's online algorithm), the empirical values $\hat{F}_{(12)}(\mathbf{w})$ do not converge uniformly to their expectations, and empirical minimization is not guaranteed to solve the problem!

4.3 Unique Empirical Minimizer Does Not Converge to Population Optimum

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, and far from the origin. Consider:

$$f_{(13)}(\mathbf{w}; (\mathbf{x}, \boldsymbol{\alpha})) = f_{(12)}(\mathbf{w}; (\mathbf{x}, \boldsymbol{\alpha})) + \epsilon \sum_i 2^{-i}(w[i]-1)^2 \tag{13}$$

where $\epsilon = 0.01$. The objective is still convex and $(1+\epsilon)$-Lipschitz. Furthermore, since the additional term is strictly convex, we have that $f_{(13)}(\mathbf{w}; z)$ is strictly convex w.r.t. $\mathbf{w}$ and so the empirical minimizer is unique.

Consider the same distribution over Z: $\mathbf{X} = 0$ while $\boldsymbol{\alpha}[i]$ are i.i.d. uniform zero or one. The empirical minimizer is the minimizer of $\hat{F}_{(13)}(\mathbf{w})$ subject to the constraints $\|\mathbf{w}\| \leq 1$. Identifying the solution to this constrained optimization problem is tricky, but fortunately not necessary. It is enough to show that the optimum of the *unconstrained* optimization problem $\mathbf{w}^*_{\mathrm{UC}} = \arg\min \hat{F}_{(13)}(\mathbf{w})$ (without constraining $\mathbf{w} \in \mathcal{W}$) has norm $\|\mathbf{w}^*_{\mathrm{UC}}\| \geq 1$. Notice that in the unconstrained problem, whenever $\boldsymbol{\alpha}_i[j] = 0$ for all $i = 1..n$, only the second term of $f_{(13)}$ depends on $\mathbf{w}[j]$ and we have $\mathbf{w}^*_{\mathrm{UC}}[j] = 1$. Since this happens a.s. for some coordinate j, we can conclude that the solution to the constrained optimization problem lies on the boundary of $\mathcal{W}$, i.e. has $\|\hat{\mathbf{w}}\| = 1$. But for such a solution we have $F_{(13)}(\hat{\mathbf{w}}) \geq \mathbb{E}_{\boldsymbol{\alpha}}\left[\sqrt{\sum_i \boldsymbol{\alpha}[i]\hat{\mathbf{w}}^2[i]}\right] \geq \mathbb{E}_{\boldsymbol{\alpha}}\left[\sum_i \boldsymbol{\alpha}[i]\hat{\mathbf{w}}^2[i]\right] = \frac{1}{2}\|\hat{\mathbf{w}}\|^2 = \frac{1}{2}$, while $F(\mathbf{w}^\star) \leq F(0) = \epsilon$.

In conclusion, no matter how big the sample size is, the unique empirical minimizer $\hat{\mathbf{w}}$ of the stochastic convex optimization problem (13) is a.s. much worse than the population optimum, $F(\hat{\mathbf{w}}) \geq \frac{1}{2} > \epsilon \geq F(\mathbf{w}^\star)$, and certainly does not converge to it.

5 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 $f(\mathbf{w}; z)$ 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 $F(\mathbf{w})$.

5.1 Empirical Minimizer converges to Population Optimum

Theorem 6. *Consider a stochastic convex optimization problem such that $f(\mathbf{w}; z)$ is λ-strongly convex and L-Lipschitz with respect to $\mathbf{w} \in \mathcal{W}$. Let $z_1, \ldots, z_n$ be an i.i.d. sample and let $\hat{\mathbf{w}}$ be the empirical minimizer. Then, with probability at least $1-\delta$ over the sample we have*

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star) \leq \frac{4L^2}{\delta\,\lambda\,n}\,. \tag{14}$$

Proof. To prove the Theorem, we use a stability argument introduced by Bousquet and Elisseeff [BE02]. Denote

$$\hat{F}^{(i)}(\mathbf{w}) = \frac{1}{n}\left(\sum_{j\neq i} f(\mathbf{w}, z_i) + f(\mathbf{w}, z'_i)\right)$$

the empirical average with z_i replaced by an independently and identically drawn z'_i, and consider its minimizer:

$$\hat{\mathbf{w}}^{(i)} = \arg\min_{\mathbf{w}\in\mathcal{W}} \hat{F}^{(i)}(\mathbf{w}).$$

We first use strong convexity and Lipschitz-continuity to establish that empirical minimization is stable in the following sense:

$$\forall z_1, \ldots, z_n, z'_i, z \in \mathcal{Z} \quad \left|f(\hat{\mathbf{w}}, z) - f(\hat{\mathbf{w}}^{(i)}, z)\right| \leq \beta_n \tag{15}$$

with $\beta_n = \frac{4L^2}{\lambda n}$ (this is referred to as "CV (Replacement) Stability" [RMP05] and is similar to "uniform stability" [BE02]). We then show that (15) implies convergence of $F(\hat{\mathbf{w}})$ to $F(\mathbf{w}^\star)$.

Claim 6.1. *Under the conditions of Theorem 6, the stability bound* (15) *holds with $\beta_n = \frac{4L^2}{\lambda n}$.*

Proof of Claim 6.1: We first calculate:

$$\begin{aligned}
&\hat{F}(\hat{\mathbf{w}}^{(i)}) - \hat{F}(\hat{\mathbf{w}}) \\
&= \frac{f(\hat{\mathbf{w}}^{(i)}, z_i) - f(\hat{\mathbf{w}}, z_i)}{n} + \frac{\sum_{j\neq i}\left(f(\hat{\mathbf{w}}^{(i)}, z_i) - f(\hat{\mathbf{w}}, z_i)\right)}{n} \\
&= \frac{f(\hat{\mathbf{w}}^{(i)}, z_i) - f(\hat{\mathbf{w}}, z_i)}{n} + \frac{f(\hat{\mathbf{w}}, z'_i) - f(\hat{\mathbf{w}}^{(i)}, z'_i)}{n} \\
&\quad + \left(\hat{F}^{(i)}(\hat{\mathbf{w}}^{(i)}) - \hat{F}^{(i)}(\hat{\mathbf{w}})\right) \qquad (16) \\
&\leq \frac{|f(\hat{\mathbf{w}}^{(i)}, z_i) - f(\hat{\mathbf{w}}, z_i)|}{n} + \frac{|f(\hat{\mathbf{w}}, z'_i) - f(\hat{\mathbf{w}}^{(i)}, z'_i)|}{n} \\
&\leq \frac{2L}{n}\left\|\hat{\mathbf{w}}^{(i)} - \hat{\mathbf{w}}\right\| \qquad (17)
\end{aligned}$$

where the first inequality follows from the fact that $\hat{\mathbf{w}}^{(i)}$ is the minimizer of $\hat{F}^{(i)}(\mathbf{w})$ and for the second inequality we use Lipschitz continuity. 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

$$\hat{F}(\hat{\mathbf{w}}^{(i)}) \geq \hat{F}(\hat{\mathbf{w}}) + \tfrac{\lambda}{2}\left\|\hat{\mathbf{w}}^{(i)} - \hat{\mathbf{w}}\right\|^2. \tag{18}$$

Combining (18) with (17) we get $\left\|\hat{\mathbf{w}}^{(i)} - \hat{\mathbf{w}}\right\| \leq 4L/(\lambda n)$. Finally from Lipschitz continuity, for any $z \in \mathcal{Z}$:

$$\left|f(\hat{\mathbf{w}}, z) - f(\hat{\mathbf{w}}^{(i)}, z)\right| \leq \tfrac{4L^2}{\lambda n} \tag{19}$$

Claim 6.2. *If the stability bound* (15) *holds, then for any* $\delta > 0$*, with probability* $1 - \delta$ *over the sample,*

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star) \leq \frac{\beta_n}{\delta}. \tag{20}$$

A similar result that is not specific to $\hat{\mathbf{w}}$, but yields only a $\sqrt{\beta_n + \frac{1}{n}}$ rate appears in [RMP05, Theorem 4.4]. The faster rate is important for us here.

Proof of Claim 6.2: Since the samples with z_i and with z_i' are identically distributed, and z_i is independent of z_i', we have:

$$\mathbb{E}\left[F(\hat{\mathbf{w}})\right] = \mathbb{E}\left[F(\hat{\mathbf{w}}^{(i)})\right] = \mathbb{E}\left[f(\hat{\mathbf{w}}^{(i)}; z_i)\right]$$

where the expectation is over $z_1, \ldots, z_n, z_i'$. This holds for all i, and so we can also write:

$$\mathbb{E}\left[F(\hat{\mathbf{w}})\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[f(\hat{\mathbf{w}}^{(i)}; z_i)\right]. \tag{21}$$

We also have:

$$\mathbb{E}\left[\hat{F}(\hat{\mathbf{w}})\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(\hat{\mathbf{w}}; z_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[f(\hat{\mathbf{w}}; z_i)\right] \tag{22}$$

Combining (21) and (22) and using (15) yields[2]:

$$\mathbb{E}\left[F(\hat{\mathbf{w}}) - \hat{F}(\hat{\mathbf{w}})\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[f(\hat{\mathbf{w}}^{(i)}, z_i) - f(\hat{\mathbf{w}}; z_i)\right] \leq \beta_n$$

We also have that $\mathbb{E}\left[F(\mathbf{w}^\star)\right] = \mathbb{E}\left[\hat{F}(\mathbf{w}^\star)\right] \geq \mathbb{E}\left[\hat{F}(\hat{\mathbf{w}})\right]$, where the equality is just equating an expectation to an expectation of an average, and the inequality follows from optimality of $\hat{\mathbf{w}}$. We can therefore conclude:

$$\mathbb{E}\left[F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star)\right] \leq \mathbb{E}\left[F(\hat{\mathbf{w}}) - \hat{F}(\hat{\mathbf{w}})\right] \leq \beta_n. \tag{23}$$

Using Markov's inequality yields (20). □

We suspect that the dependence on δ in the above bound can be improved to $\log(1/\delta)$, matching the dependence on δ in the online-to-batch guarantee (8) and the gurantees for the generalized linear case. For more details, see Appendix A.

5.2 But Without Uniform Convergence!

We now turn to ask whether the convergence of the empirical minimizer in this case is a result of uniform convergence.

Consider augmenting the objective function $f_{(12)}$ of Section 4 with a strongly convex term:

$$f_{(24)}(\mathbf{w}; \mathbf{x}, \boldsymbol{\alpha}) = f_{(12)}(\mathbf{w}; \mathbf{x}, \boldsymbol{\alpha}) + \frac{\lambda}{2}\left\|\mathbf{w}\right\|^2. \tag{24}$$

The modified objective $f_{(24)}(\cdot;\cdot)$ is λ-strongly convex and $(1+\lambda)$-Lipschitz over the domain $\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}\| \leq 1\}$ and thus satisfies the conditions of Theorem 6.

Consider the same distribution over $Z = (\mathbf{X}, \boldsymbol{\alpha})$ used in Section 4: $\mathbf{X} = 0$ and $\boldsymbol{\alpha}$ is an i.i.d. sequence of uniform zero/one Bernoulli variables. Recall that almost surely we have a coordinate j that is never "observed", i.e. such that $\forall_i \boldsymbol{\alpha}_i[j] = 0$. Consider a vector $t\mathbf{e}_j$ of magnitude $0 < t \leq 1$ in the direction of this coordinate. We have that $\hat{F}_{(24)}(t\mathbf{e}_j) = \frac{\lambda}{2}t^2$ but $F_{(24)}(t\mathbf{e}_j) = \frac{1}{2}t + \frac{\lambda}{2}t^2$. Hence $F_{(24)}(t\mathbf{e}_j) - \hat{F}_{(24)}(t\mathbf{e}_j) = t/2$.

In particular, we can set $t = 1$ and establish $\sup_{\mathbf{w}\in\mathcal{W}}(F_{(24)}(\mathbf{w}) - \hat{F}_{(24)}(\mathbf{w})) \geq \frac{1}{2}$ regardless of the sample size. We see then that the empirical averages $\hat{F}_{(24)}(\mathbf{w})$ do *not* converge uniformly to their expectations, even as the sample size increases.

5.3 Not Even Local Uniform Convergence

For any $\epsilon > 0$, consider limiting our attention only to predictors that are close to being population optimal:

$$\mathcal{W}_\epsilon = \left\{\mathbf{w} \in \mathcal{W} : F_{(24)}(\mathbf{w}) \leq F_{(24)}(\mathbf{w}^\star) + \epsilon\right\}.$$

Setting $t = \epsilon$ we have $t\mathbf{e}_j \in \mathcal{W}_\epsilon$ (focusing for convenience on $\lambda < 1$) and so:

$$\sup_{\mathbf{w}\in\mathcal{W}_\epsilon} (F_{(24)}(\mathbf{w}) - \hat{F}_{(24)}(\mathbf{w})) \geq \tfrac{\lambda}{2}\epsilon^2 \tag{25}$$

regardless of the sample size. And so, even in an arbitrarily small neighborhood of the optimum, the empirical values $\hat{F}_{(24)}(\mathbf{w})$ do *not* converge uniformly to their expected values even as $n \to \infty$. This is in sharp contrast to essentially all other results on stochastic optimization and learning that we are aware of.

5.4 Bounding Population Sub-Optimality in term of Empirical Sub-Optimality

A practical question related to uniform convergence is whether we can obtain a uniform bound on the population sub-optimality in terms of the empirical sub-optimality, as in Theorem 2. 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 6, that with probability at least $1 - \delta$, for all $\mathbf{w} \in \mathcal{W}$:

$$F(\mathbf{w}) - F(\mathbf{w}^\star) \leq \sqrt{\frac{2L^2}{\lambda}}\sqrt{\hat{F}(\mathbf{w}) - \hat{F}(\hat{\mathbf{w}})} + \frac{4L^2}{\delta\,\lambda\,n} \tag{26}$$

It is important to emphasize that this is an immediate consequence of (14) and does not involve any further stochastic properties of $\hat{F}$ nor 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 more favorable linear dependence of Theorem 2. Unfortunately, as we show next, this is the best that can be ensured.

Consider the objective $f_{(24)}$ and the same distribution over $Z = (\mathbf{X}, \boldsymbol{\alpha})$ discussed above and recall that $t\mathbf{e}_j$ is a vector of magnitude t along a coordinate j s.t. $\forall_i \boldsymbol{\alpha}_i[j] = 0$. We have that $\hat{F}_{(24)}(t\mathbf{e}_j) - \hat{F}_{(24)}(\hat{\mathbf{w}}) = \frac{\lambda}{2}t^2$ and so setting $t = \sqrt{2\epsilon/\lambda}$, we get an ϵ-empirical-suboptimal vector with population sub-optimality $F_{(24)}(t\mathbf{e}_j) - F_{(24)}(0) = \frac{1}{2}t + \frac{\lambda}{2}t^2 = \sqrt{\frac{\epsilon}{2\lambda}} + \epsilon$.

[2] This is a modification of a derivation extracted from the proof of Theorem 12 in [BE02]

This establishes that the dependence on $\sqrt{\frac{\epsilon}{\lambda}}$ in the first term of (26) is tight, and the situation is qualitatively different than the generalized linear case.

5.5 Contradiction to Vapnik?

At this point, a reader familiar with Vapnik's work on necessary and sufficient conditions for consistency of empirical minimization (i.e. conditions for $F(\hat{\mathbf{w}}) \to F(\mathbf{w}^\star)$) might be confused.

In seeking such necessary and sufficient conditions [Vap98, Chapter 3], Vapnik excludes certain consistent settings where the consistency is so-called "trivial". The main example of an excluded setting is one in which there is one hypothesis $\mathbf{w}^0$ that dominates all others, i.e. $f(\mathbf{w}^0; z) < f(\mathbf{w}; z)$ for all $\mathbf{w} \in \mathcal{W}$ and all $z \in \mathcal{Z}$ [Vap98, Figure 3.2]. When this is the case, empirical minimization will be consistent regardless of the behavior of $\hat{F}(\mathbf{w})$ for $\mathbf{w} \neq \mathbf{w}^0$.

In order to exclude such "trivial" cases Vapnik defines strict (aka "non-trivial") consistency of empirical minimization as (in our notation):

$$\inf_{F(\mathbf{w}) \geq c} \hat{F}(\mathbf{w}) \xrightarrow{P} \inf_{F(\mathbf{w}) \geq c} F(\mathbf{w}) \quad \geq \quad c \tag{27}$$

for all $c \in \mathbb{R}$, where the convergence is in probability. This condition indeed ensures that $F(\hat{\mathbf{w}}) \xrightarrow{P} F(\mathbf{w}^\star)$. Vapnik's Key Theorem on Learning Theory [Vap98, Theorem 3.1] then states that *strict* consistency of empirical minimization is equivalent to one-sided uniform convergence. "One-sided" meaning requiring only $\sup(F_{(24)}(\mathbf{w}) - \hat{F}_{(24)}(\mathbf{w})) \xrightarrow{P} 0$, rather then $\sup \left| F_{(24)}(\mathbf{w}) - \hat{F}_{(24)}(\mathbf{w}) \right| \xrightarrow{P} 0$. Note that the analysis above shows the lack of such one-sided uniform convergence.

In the example presented above, even though Theorem 6 establishes $F(\hat{\mathbf{w}}) \xrightarrow{P} F(\mathbf{w}^\star)$, the consistency isn't "strict" by the definition above. To see this, for any $c > 0$, consider the vector $t\mathbf{e}_j$ (where $\forall_i \boldsymbol{\alpha}_i[j] = 0$) with $t = 2c$. We have $F(t\mathbf{e}_j) = \frac{1}{2}t + \frac{\lambda}{2}t^2 > c$ but $\hat{F}_{(24)}(t\mathbf{e}_j) = \frac{\lambda}{2}t^2 = 2\lambda c^2$. Focusing on $\lambda = \frac{1}{2}$ we get:

$$\inf_{F(\mathbf{w}) \geq c} \hat{F}(\mathbf{w}) \leq c^2 \tag{28}$$

almost surely for any sample size n, violating the strict consistency requirement (27). The fact that the right-hand-side of (28) is strictly greater then $F(\mathbf{w}^\star) = 0$ is enough for obtaining (non strict) consistency of empirical minimization, but this is not enough for satisfying strict consistency.

We emphasize that stochastic convex optimization is far from "trivial" in that there is no dominating hypothesis that will always be selected. Although for convenience of analysis we took $\mathbf{X} = 0$, one should think of situations in which $\mathbf{X}$ is stochastic with an unknown distribution.

We see then that there is no mathematical contradiction here to Vapnik's Key Theorem. Rather, we see a demonstration that strict consistency is too strict a requirement, and that interesting, non-trivial, learning problems might admit non-strict consistency which is *not* equivalent to one-sided uniform convergence. We see that uniform convergence is a sufficient, but not at all necessary, condition for consistency of empirical minimization in non-trivial settings.

6 Regularization

We now return to the case where $f(\mathbf{w}, z)$ 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

$$\hat{\mathbf{w}}_\lambda = \min_{\mathbf{w} \in \mathcal{W}} \left(\tfrac{\lambda}{2} \|\mathbf{w}\|^2 + \tfrac{1}{n} \sum_{i=1}^{n} f(\mathbf{w}, z_i) \right) , \tag{29}$$

where λ is a parameter that will be determined later. The following theorem establishes that the minimizer of (29) is a good solution to the stochastic convex optimization problem:

Theorem 7. *Let $f : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ be such that $\mathcal{W}$ is bounded by B and $f(\mathbf{w}, z)$ is convex and L-Lipschitz with respect to $\mathbf{w}$. Let $z_1, \ldots, z_n$ be an i.i.d. sample and let $\hat{\mathbf{w}}_\lambda$ be the minimizer of* (29) *with $\lambda = \sqrt{\frac{16L^2}{\delta B^2 n}}$. Then, with probability at least $1 - \delta$ we have*

$$F(\hat{\mathbf{w}}_\lambda) - F(\mathbf{w}^\star) \;\leq\; 4\sqrt{\frac{L^2 B^2}{\delta n}} \left(1 + \frac{8}{\delta n} \right) .$$

Proof. Let $r(\mathbf{w}; z) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + f(\mathbf{w}; z)$ and let $R(\mathbf{w}) = \mathbb{E}_z [r(\mathbf{w}, z)]$. Note that $\hat{\mathbf{w}}_\lambda$ is the empirical minimizer for the stochastic optimization problem defined by $r(\mathbf{w}; z)$.

We apply Theorem 6 to $r(\mathbf{w}; z)$, to this end note that since f is L-Lipschitz and $\forall \mathbf{w} \in \mathcal{W}$, $\|\mathbf{w}\| \leq B$ we see that r is in fact $L + \lambda B$-Lipschitz. Applying Theorem 6 now we see that

$$\tfrac{\lambda}{2} \|\hat{\mathbf{w}}_\lambda\|^2 + F(\hat{\mathbf{w}}_\lambda) = R(\hat{\mathbf{w}}_\lambda) \leq \inf_{\mathbf{w}} R(\mathbf{w}) + \frac{4(L + \lambda B)^2}{\delta \lambda n}$$
$$\leq R(\mathbf{w}^\star) + \frac{4(L + \lambda B)^2}{\delta \lambda n} = \tfrac{\lambda}{2} \|\mathbf{w}^\star\|^2 + F(\mathbf{w}^\star) + \frac{4(L + \lambda B)^2}{\delta \lambda n}$$

Now note that $\|\mathbf{w}^\star\| \leq B$ and so we get that

$$F(\hat{\mathbf{w}}_\lambda) \leq F(\mathbf{w}^\star) + \frac{\lambda}{2} B^2 + \frac{4(L + \lambda B)^2}{\delta \lambda n}$$
$$\leq F(\mathbf{w}^\star) + \frac{\lambda}{2} B^2 + \frac{8L^2}{\delta \lambda n} + \frac{8\lambda B^2}{\delta n}$$

Plugging in the value of λ given in the theorem statement we see that

$$F(\hat{\mathbf{w}}_\lambda) \leq F(\mathbf{w}^\star) + 4\sqrt{\frac{L^2 B^2}{\delta n}} + \frac{32}{\delta n} \sqrt{\frac{L^2 B^2}{\delta n}}$$

This gives us the required bound. □

From the above theorem and the discussion in Section 4 we see that regularization is essential for convex stochastic optimization. It is interesting to contrast this with the online learning algorithm of Zinkevich [Zin03]. 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 [Sha07] showed that Zinkevich's online learning

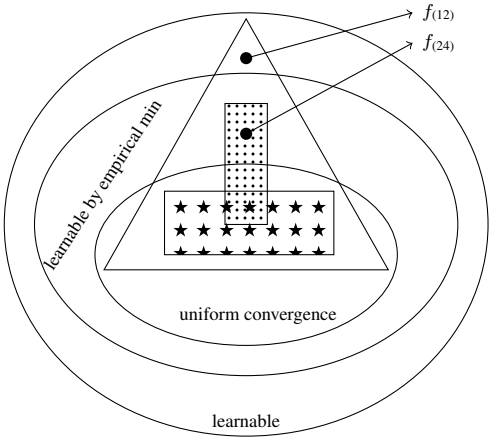

Figure 1: Lipschitz-continuous convex problems (triangle) are all learnable, but not necessarily using empirical minimization. Lipschitz-continuous strongly convex problems (dotted rectangle) are all learnable with empirical minimization, but uniform convergence might not hold. For bounded generalized linear problems (starred rectangle), uniform convergence always holds. Our two separating examples are also indicated.

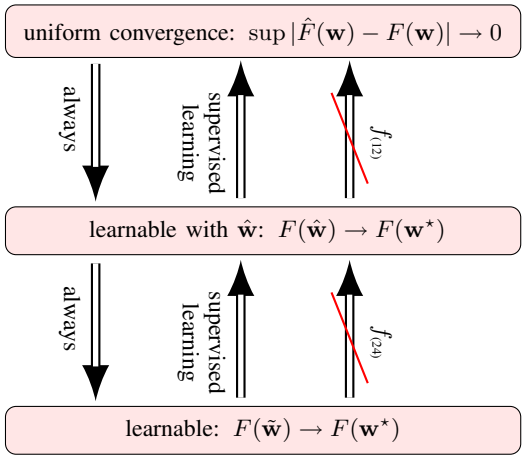

Figure 2: Relationship between different properties of stochastic optimization problems.

algorithm can be viewed as approximate coordinate ascent optimization of the dual of the regularized problem (29). 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 (29) on $z_1, \ldots, z_{i-1}$. The key, then, seems to be regularization, rather then a procedural online versus global minimization approach.

6.1 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 $f_{(12)}$ of Section 4. Even over a restricted domain $\mathcal{W}_r = \{\mathbf{w} : \|\mathbf{w}\| \leq r\}$, for arbitrarily small $r > 0$, the empirical averages $\hat{F}(\mathbf{w})$ do *not* uniformly converge to $F(\mathbf{w})$ and $\Pr\left(\limsup_{n\to\infty} \sup_{\mathbf{w}\in\mathcal{W}_r} \left|\hat{F}(\mathbf{w}) - F(\mathbf{w})\right| > 0\right) = 1$. Furthermore, consider replacing the regularization term $\lambda\|\mathbf{w}\|^2$ with a constraint on the norm of $\|\mathbf{w}\|$, namely, solving the problem

$$\tilde{\mathbf{w}}_r = \arg\min_{\|\mathbf{w}\|\leq r} \hat{F}(\mathbf{w}) \tag{30}$$

As we show below, we cannot solve the stochastic optimization problem by setting r in a distribution-independent way (i.e. without knowing the solution...). To see this, note that when $\mathbf{X} = 0$ a.s. we must have $r \to 0$ to ensure $F(\tilde{\mathbf{w}}_w) \to F(\mathbf{w}^\star)$. However, if $\mathbf{X} = \mathbf{e}_1$ a.s., we must set $r \to 1$. No constraint will work for all distributions over $\mathbb{Z} = (\mathbf{X}, \boldsymbol{\alpha})$! 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 parameter such as λ, and adding a regularization term of the form $\frac{\lambda}{2}\|\mathbf{w}\|^2$ is viewed as a proxy for bounding the norm $\|\mathbf{w}\|$.

7 Summary

Following the work of Zinkevich [Zin03], 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 key role and ultimately did reach an understanding of stochastic convex optimization that does not rely on online techniques (Figure 1).

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 learnable, it is learnable using empirical minimization [ABCH97]. Many might think that this principal, namely that a problem is learnable iff it is learnable using empirical minimization, extends also the "General Setting of Learning" [Vap95] which includes also the stochastic convex optimization problem studied here.

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 reasonable loss functions, in order to make $f(\mathbf{w}; \mathbf{x}, y) = \ell(\text{pred}(\mathbf{w}, \mathbf{x}), y)$ convex for both posi-

tive and negative labels, we must essentially make the prediction function $\mathrm{pred}(\mathbf{w}, \mathbf{x})$ both convex and concave in $\mathbf{w}$, i.e. linear. And so stochastic (or online) convex optimization problems that correspond to supervised problems are often generalized linear problems.

To summarize, although there is no contradiction to the work of Vapnik [Vap95] or of Alon *et al* [ABCH97], we see that learning in the General Setting is more complex than we perhaps appreciate. Empirical minimization might be consistent without local uniform convergence, and more surprisingly, learning might be possible, but not by empirical minimization (Figure 2).

Acknowledgments

We would like to thank Leon Bottou, Tong Zhang, and Vladimir Vapnik for helpful discussions.

References

[ABCH97] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997.

[BE02] O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.

[CCG04] 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.

[HKKA06] 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.

[HKLW91] David Haussler, Michael Kearns, Nick Littlestone, and Manfred K. Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*, 95(2):129–161, December 1991.

[KT08] S.M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *NIPS*, 2008.

[Pol84] D. Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.

[RMP05] S. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3(4):397–419, 2005.

[Sha07] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

[SSS08] K. Sridharan, N. Srebro, and S. Shalev-Shwartz. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 22*, 2008.

[Vap95] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[Vap98] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[VG05] J.L. Verger-Gaugry. Covering a ball with smaller equal balls in $\mathbb{R}^n$. *Discrete Comput. Geom.*, 33(1):143–155, 2005.

[vLB04] U. von Luxburg and O. Bousquet. Distance–based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5:669–695, 2004.

[Zin03] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

A High Confidence Bounds

The bounds in Theorems 6 and 7 have polynomial rather than logarithmic dependence on the confidence parameter δ. This leads to the question of whether these bounds can be improved to depend just on $\log(1/\delta)$, matching the dependence on δ in the online-to-batch guarantees (7) and (8). While we suspect this might be the case, the question remains open.

We emphasize that the question here pertains to the bound on the convergence *of the empirical minimizer*. The online-to-batch guarantees apply only to a specific procedurally defined predictor, which is not the empirical minimizer. Another simple way to achieve a logarithmic dependence on $1/\delta$ is to use empirical minimization combined with a generic boosting-the-confidence method [HKLW91]. Again, this leads to a high-confidence bound for a different learning rule, based on the empirical minimizer, but is not the empirical minimizer.

As for results regarding the empirical minimizer itself, we note that it is possible to get high-confidence bounds, with only a logarithmic dependence on $1/\delta$. However, these bounds come at the price of worse dependence on the other parameters of the learning problem. For instance, if $F(\mathbf{w})$ is twice continuously differentiable, with a uniform upper bound $\lambda_{\max}$ on the eigenvalues of its Hessian, and the conditions of Theorem 6 hold, we get that with probability at least $1-\delta$:

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star) \;\leq\; \mathcal{O}\left(\frac{L^2 \log(1/\delta)\lambda_{\max}}{\lambda^2 n}\right). \qquad (31)$$

Also, under the conditions of Theorem 6 and without any additional assumption, Bousquet and Elisseeff [BE02] provide arguments for a bound of the form

$$F(\hat{\mathbf{w}}) - F(\mathbf{w}^\star) \;\leq\; \mathcal{O}\left(\frac{L^2}{\lambda}\sqrt{\frac{\log(1/\delta)}{n}}\right). \qquad (32)$$

Unfortunately, neither of these two bounds is sufficient for obtaining a version of Theorem 7 which matches the online-to-batch guarantee (8) or the bound of Theorem 1 for the generalized linear case. Optimizing for the value of λ as a function of the sample size, we get that the bound on the unregularized objective function in Theorem 7 is replaced by

$$F(\hat{\mathbf{w}}_\lambda) - F(\mathbf{w}^\star) \leq \mathcal{O}\left(\left(\frac{B^4 L^2 \log(1/\delta)\lambda_{\max}}{n}\right)^{1/3}\right)$$

if we use (31), or

$$F(\hat{\mathbf{w}}_\lambda) - F(\mathbf{w}^\star) \leq \mathcal{O}\left(\left(\frac{B^4 L^4 \log(1/\delta)}{n}\right)^{1/4}\right)$$

if we use (32). In particular, the dependence on the sample size n is significantly worse than $n^{-1/2}$.