Proceedings of Machine Learning Research vol 178:1–20, 2022 35th Annual Conference on Learning Theory

Uniform Stability for First-Order Empirical Risk Minimization

Amit Attia AMITATTIA@MAIL.TAU.AC.IL
Blavatnik School of Computer Science, Tel Aviv University

Tomer Koren TKOREN@TAUEX.TAU.AC.IL
Blavatnik School of Computer Science, Tel Aviv University, and Google Research

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We consider the problem of designing uniformly stable first-order optimization algorithms for empirical risk minimization. Uniform stability is often used to obtain generalization error bounds for optimization algorithms, and we are interested in a general approach to achieve it. For Euclidean geometry, we suggest a black-box conversion which given a smooth optimization algorithm, produces a uniformly stable version of the algorithm while maintaining its convergence rate up to logarithmic factors. Using this reduction we obtain a (nearly) optimal algorithm for smooth optimization with convergence rate $\widetilde{O}(1/T^2)$ and uniform stability $O(T^2/n)$, resolving an open problem of Chen et al. (2018); Attia and Koren (2021). For more general geometries, we develop a variant of Mirror Descent for smooth optimization with convergence rate $\widetilde{O}(1/T)$ and uniform stability $O(T/n)$, leaving open the question of devising a general conversion method as in the Euclidean case.

1. Introduction

We consider a canonical problem in machine learning: empirical risk minimization using first-order convex optimization. Given a training sample $S = (z_1, \ldots, z_n)$ of n instances, the goal is to minimize the empirical risk $L_S(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(x; z_i)$ where $\ell(\cdot, z)$ is a convex loss function. Our focus is on the smooth case which contains a variety of first-order algorithms including gradient descent (GD) and Nesterov's celebrated accelerated gradient method (Nesterov, 1983). In statistical learning, the empirical risk is used as a proxy and our true goal is to minimize the *population risk* $L(x) \triangleq \mathbb{E}_{z\in\mathcal{D}}[\ell(x; z)]$ of an unknown distribution $\mathcal{D}$. The performance of a learning algorithm $\mathcal{A}$ is evaluated by the expected *excess population risk*,

$$\mathbb{E}_S[L(A(S)) - L(x^\star)],$$

where $x^\star \in \arg\min L(x)$, which is often bounded by managing the trade-off between the expected optimization and generalization errors:

$$\mathbb{E}_S[L(A(S)) - L(x^\star)] = \mathbb{E}_S[\underbrace{L(A(S)) - L_S(A(S))}_{\text{generalization}}] + \mathbb{E}_S[\underbrace{L_S(A(S)) - L_S(x^\star)}_{\text{optimization}}].$$

Thus, for a given optimization method to minimize the empirical risk, we are often interested in bounding the generalization error of its solution. A fundamental framework for obtaining such bounds is algorithmic stability (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2009).

Algorithmic stability has emanated as a central tool for generalization analysis of learning algorithms. The pioneering work of Bousquet and Elisseeff (2002) introduced the notion of *uniform* *stability*, arguably the most common notion of algorithmic stability in learning theory. Essentially, to this day, stability analysis is the only general approach for obtaining tight, dimension free generalization bounds for convex optimization algorithms applied to the empirical risk (see Shalev-Shwartz et al., 2009; Feldman, 2016).

Although stability analysis has proved to be an effective tool for generalization bounds, it unfortunately applies for specific combinations of algorithms, objectives and geometries. For example, gradient descent is known to be uniformly stable for smooth objectives in ℓ_2 geometry (Hardt et al., 2016; Feldman and Vondrak, 2018; Chen et al., 2018), but the stability analysis fails in other geometries (e.g., ℓ_1) due to a lack of the contractivity property for gradient steps (Asi et al., 2021). Another example is Nesterov's gradient method, for which uniform stability is quadratic in the number of steps over quadratic objectives (Chen et al., 2018) but for general smooth objectives grows exponentially fast (Attia and Koren, 2021).

The last example is particularly intriguing: as highlighted by Chen et al. (2018) and Attia and Koren (2021), it is currently not known whether there exists an optimal method for smooth optimization, with convergence rate $O(1/T^2)$, which also exhibits the optimal uniform stability rate $O(T^2/n)$ for general smooth objectives. This is an important issue as momentum-based methods, inspired by Nesterov's optimal method, are being extensively used for empirical risk minimization in practice and understanding their stability properties would help in shedding light on the generalization ability of such methods. Developing an optimal and uniformly stable method (or proving that one does not exist) was thus left as an open problem by Chen et al. (2018); Attia and Koren (2021).

1.1. Contributions

In this paper, motivated by the open question of Chen et al. (2018); Attia and Koren (2021), we study general techniques for uniformly stable empirical risk optimization with smooth and convex objectives. First, we focus on the Euclidean case and give a general and widely-applicable technique for converting optimization algorithms to uniformly stable ones. Then, we move on to develop uniformly stable algorithms for smooth convex optimization in more general normed spaces.

General reduction in Euclidean geometry. We provide an algorithm, USOL2 (see Algorithm 1), which performs a black-box conversion from a given optimization algorithm for convex, smooth and Lipschitz objectives to a uniformly stable algorithm with nearly the same convergence rate. Following is an informal version of our first main result (stated formally at Theorem 6).

Theorem 1 (informal). *Assume an optimization algorithm $\mathcal{A}$ with convergence rate $O(T^{-\gamma})$ over convex, smooth and Lipschitz functions w.r.t. the Euclidean norm. Then applying USOL2 to $\mathcal{A}$ yields an algorithm with convergence rate $\widetilde{O}(T^{-\gamma})$ whose T'th iterate is $O(T^{\gamma}/n)$-uniformly stable.*

Thus, the conversion preserves the rate of $\mathcal{A}$ (up to logarithmic factors) and exhibits essentially the best convergence vs. stability trade-off one could hope for: indeed, any improvement to one of the rates (without compromising the other) would lead to a contradiction to statistical lower bounds (this is discussed is detail by Chen et al., 2018 [1]). Applying this result to Nesterov's accelerated gradient method, we obtain an algorithm with (nearly) *optimal* convergence rate $\widetilde{O}(1/T^2)$ and a

1. We remark that the setting of Chen et al. (2018) did not include a Lipschitz assumption (in addition to smoothness), as they discuss in their Section 4.2. That said, a straightforward modification of the proof of Theorem 7 in Chen et al. (2018) can accommodate the Lipschitz assumption by a simple scaling of the loss function.

matching *optimal* stability rate of $O(T^2/n)$, resolving an open problem posed by Chen et al. (2018) and reiterated by Agarwal et al. (2020) and Attia and Koren (2021).

Our conversion procedure is based on two simple observations. The first is that the minimizer of a strongly convex objective is uniformly stable. The second is that with smoothness, converging very close to the minimizer of a regularized objective comes almost for free (at the cost of only a logarithmic factor) since strong convexity and smoothness together allow for linear rates of convergence. Carefully combining the two leads to a simple yet effective way to achieve stability in smooth convex optimization with a minimal degradation in convergence rate.

Stable Mirror Descent for general norms. We move on to address uniform stability in more general normed spaces. A general approach to optimization with general norms is the so called Mirror Descent (Nemirovskij and Yudin, 1983) which has convergence rate $O(1/T)$ for smooth objectives (e.g., Bubeck, 2015) and accelerated variants with rate $O(1/T^2)$ (Tseng, 2008; Allen-Zhu and Orecchia, 2017). A natural followup question to our investigation in the Euclidean case is whether there exists a variant of (accelerated) Mirror Descent with a similar convergence rate, which is also uniformly stable.

This general scenario poses additional challenges, as even for simple non-accelerated Mirror Descent, previous work by Asi et al. (2021) gave an indication that a standard Mirror Descent gradient step fails to be contractive (in fact, it is slightly expansive). This questions the primary approach for proving stability of iterative methods in the context of Mirror Descent.

As it turns out, a general conversion scheme as the one we use in Algorithm 1 does not easily extends to general geometries. In a nutshell, the issue is the following: given a black-box algorithm for smooth and convex optimization, the standard reduction for obtaining a linear rate assuming the function is also strongly convex is based on a contraction argument relating the convergence rate upper bound to the squared distance from the minimizer, via strong convexity; for general norms, however, the convergence (e.g., of Mirror Descent) depends in general on the Bregman divergence rather than the squared distance and the argument does not go through.

Instead, we devise a specialized algorithm called USMD (Algorithm 2), which obtains the following result (stated formally at Theorem 13), and leave the problem of designing a generic conversion method for general geometries as an open question.

Theorem 2 (informal). *Assume a convex loss function ℓ, which is smooth and Lipschitz (w.r.t. a norm $\|\cdot\|$) and a 1-strongly convex regularization R (w.r.t. the same norm). Then USMD applied to the empirical risk with regularizer R has convergence rate $\widetilde{O}(1/T)$ and its output after T steps is $O(T/n)$-uniformly stable.*

Hence, USMD has the desired uniform stability of $O(T/n)$ and has nearly the same rate as Mirror Descent. The algorithm is also based on the simple approach of optimizing a regularized objective; however, as regularization in general norms can impair smoothness (e.g., $\frac{1}{2}\|\cdot\|_p^2$ for $1 < p < 2$ is not smooth, see Appendix C), our analysis is based on relative smooth convex optimization (Lu et al., 2018) (the regularization is smooth w.r.t. itself), exploiting the linear rate to converge to a stable minimizer.

Examples. We discuss two example applications of our general algorithm.

- *ℓ_p geometry* ($1 < p \le 2$): In the case where $\mathcal{X}$ is the ℓ_p unit ball and $\ell(\cdot, z)$ is convex, smooth and Lipschitz w.r.t. ℓ_p norm for all z, applying Theorem 13 with the mirror map $R(x) = \|x\|_p^2/2(p-1)$, USMD has convergence rate $\widetilde{O}(1/(p-1)T)$, and is $O(T/n)$-uniformly stable.

- ℓ_1 *geometry*: In the case where $\mathcal{X} = \{x \in \mathbb{R}^d_{\geq 0} : \|x\|_1 = 1\}$ and $\ell(\cdot, z)$ is convex, smooth and Lipschitz w.r.t. ℓ_1 norm for all z, applying Theorem 13 with negative entropy as the mirror map $R(x) = \sum_{i=1}^d x_i \log x_i$, the convergence rate is $\widetilde{O}(\log(d)/T)$,[2] and the algorithm is $O(T/n)$-uniformly stable.

Open problems. A couple of interesting questions remain open for investigation. The first is whether we can remove the additional log factors from both the Euclidean and general geometry methods, thus obtaining stability "for free." The second is whether we can devise a general conversion method, analogous to the one we developed in the Euclidean case, that would apply to more general norms.

1.2. Related work

Classical generalization theory appealed to uniform convergence of the empirical risk to the population risk. Without further assumptions on convex functions, the rate of uniform convergence for stochastic convex optimization is dimension-dependent and lower bounded by $\Omega(\sqrt{d/n})$ (Shalev-Shwartz et al., 2010; Feldman, 2016). Using a stability analysis, recent progress was made on stochastic optimization generalization bounds of convex risk minimizers, starting with the influential work of Bousquet and Elisseeff (2002) and Shalev-Shwartz et al. (2009). A variety of notions of algorithmic stability exists in the literature and differ in the distance measure and aggregation of multiple changes (Bousquet and Elisseeff, 2002; Mukherjee et al., 2006; Shalev-Shwartz et al., 2010; London, 2017; Lei and Ying, 2020). Data dependent generalization bounds based on stability arguments were also studied by Maurer (2017); Kuzborskij and Lampert (2018). Recent work derived tighter bounds of generalization from stability (Feldman and Vondrak, 2018, 2019; Bousquet et al., 2020; Klochkov and Zhivotovskiy, 2021), and the approach of stability analysis has been influential in a variety of settings (e.g., Koren and Levy, 2015; Gonen and Shalev-Shwartz, 2017; Charles and Papailiopoulos, 2018).

Significant interest in the stability properties of iterative methods has arisen recently. The work of Hardt et al. (2016) gave the first bounds on the uniform stability of stochastic gradient descent (SGD) for convex and smooth optimization, showing it grows linearly with the number of optimization steps. Their result apply with minor modification to full-batch gradient descent (GD) as seen in following work (Feldman and Vondrak, 2018; Chen et al., 2018). Bounds for the stability of SGD and GD in the non-smooth case was studied by Lei and Ying (2020); Bassily et al. (2020) and revealed a significant gap in stability between smooth and non-smooth optimization, indicating the importance of smoothness for stability. Furthermore, algorithmic stability was instrumental in stochastic mini-batched iterative optimization (e.g., Wang et al., 2017; Agarwal et al., 2020), and has been pivotal to the design and analysis of differentially private optimization algorithms (Wu et al., 2017; Bassily et al., 2019; Feldman et al., 2020), both of which focused mainly on smooth optimization.

2. Throughout, logarithmic factor of the space dimension are *not* suppressed by the $\widetilde{O}$ notation.

2. Preliminaries

2.1. Smooth convex optimization

In this work we are interested in optimization of convex and smooth functions over a closed convex set $\mathcal{X} \subseteq \mathbb{R}^d$. A function f is said to be β-smooth w.r.t. a norm $\|\cdot\|$ if its gradient is β-Lipschitz w.r.t. $\|\cdot\|$, namely $\|\nabla f(x) - \nabla f(y)\|_* \leq \beta\|x - y\|$ for all $x, y \in \mathcal{X}$. Here $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. This smoothness condition also yields the following quadratic bound for all $x, y \in \mathcal{X}$: $f(y) \leq f(x) + \nabla f(x) \cdot (y - x) + \frac{\beta}{2}\|y - x\|^2$. A function f is said to be α-strongly convex w.r.t. a norm $\|\cdot\|$ if for all $x, y \in \mathcal{X}$, we have $f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\alpha}{2}\|y - x\|^2$.

2.2. Algorithmic stability

In this work we consider the well known *uniform stability* (Bousquet and Elisseeff, 2002) in the following general setting of supervised learning. There is an unknown distribution $\mathcal{D}$ over a sample set $\mathcal{Z}$ from which examples are drawn. Given a training set $S = (z_1, \ldots, z_n)$ of n samples drawn i.i.d. from $\mathcal{D}$, the objective is finding a model $x \in \mathcal{X}$ with a small *population risk*:

$$L(x) \triangleq \mathbb{E}_{z\sim\mathcal{D}}[\ell(x; z)],$$

where $\ell(x; z)$ is the loss of the model described by x on an example z. We cannot evaluate the population risk directly, thus, optimization will be applied on the *empirical risk* with respect to the sample S, given by

$$L_S(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(x; z_i).$$

We use the following notion of *uniform stability*.[3]

Definition 3 (uniform stability). Algorithm A is ϵ-uniformly stable if for all $S, S' \in \mathcal{Z}^n$ such that S, S' differ in at most one example, the corresponding outputs $A(S)$ and $A(S')$ satisfy

$$\sup_{z\in\mathcal{Z}} |\ell(A(S); z) - \ell(A(S'); z)| \leq \epsilon.$$

A known result of Bousquet and Elisseeff (2002) is a bound on the expected generalization error of an ϵ-uniformly stable algorithm $\mathcal{A}$,

$$\mathbb{E}_S[L(\mathcal{A}(S)) - L_S(\mathcal{A}(S))] \leq \epsilon.$$

2.3. Relatively-smooth convex optimization

In order to handle regularization over general norms we will need the framework of relatively smooth and convex functions. Given a 1-strongly convex function $R : \mathcal{X} \mapsto \mathbb{R}$, the Bregman divergence of R is defined as $D_R(y, x) \triangleq R(y) - R(x) - \nabla R(x) \cdot (y - x)$. With the definition of Bregman divergence we can define relative smooth and strong convexity.

3. The definition is suitable for deterministic algorithms and is sufficient for the scope of this work. A more general definition exists for randomized algorithms (e.g., Hardt et al., 2016; Feldman and Vondrak, 2018).

Definition 4 (relative strong convexity). A function $f : \mathcal{X} \mapsto \mathbb{R}$ is α-strongly convex relative to $R : \mathcal{X} \mapsto \mathbb{R}$ if for any $x, y \in \mathcal{X}$, $f(x) + \nabla f(x) \cdot (y - x) + \alpha D_R(y, x) \leq f(y)$.

Definition 5 (relative smoothness). A function $f : \mathcal{X} \mapsto \mathbb{R}$ is β-smooth relative to $R : \mathcal{X} \mapsto \mathbb{R}$ if for any $x, y \in \mathcal{X}$, $f(x) + \nabla f(x) \cdot (y - x) + \beta D_R(y, x) \geq f(y)$.

Our analysis relates to the so called Mirror Descent method which is used in both smooth and relatively-smooth convex optimization (Nemirovskij and Yudin, 1983; Lu et al., 2018). The standard step of Mirror Descent over β-smooth (or relatively smooth) functions with a 1-strongly convex regularization $R(\cdot)$ is as follows:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \big\{ \nabla f(x_t) \cdot (x - x_t) + \beta D_R(x, x_t) \big\}. \tag{1}$$

3. General reduction in Euclidean geometry

In this section, we present a general method for converting a black-box smooth optimization algorithm with a given convergence guarantee to a uniformly stable algorithm with nearly the same convergence rate. Let $\mathcal{A}$ be an iterative algorithm for minimization of β-smooth convex functions with convergence rate of the form

$$f(x_t) - f(x^\star) \leq \frac{C\beta \|x_0 - x^\star\|^2}{t^\gamma},$$

where $x^\star \in \arg\min_{x \in \mathcal{X}} f(x)$, $x_t = \mathcal{A}(f, \beta, x_0, t)$ is the output of the algorithm at step t on a β-smooth function f initialized at x_0, and constants $C, \gamma > 0$. Note that it is enough to consider $0 < \gamma \leq 2$, as $O(1/t^2)$ is the optimal convergence rate for smooth optimization.

Algorithm 1: $\mathsf{USOL2}(\mathcal{A})$: Uniformly Stable Optimization in L2

Input: Sample $S = (z_i)_{i=1}^n$, base algorithm $\mathcal{A}$, initialization x_0, parameters β,G,D
Output: A sequence of iterates $\{x_t\}$.
$\alpha_0 \leftarrow \frac{\beta}{4}$; $y_0 \leftarrow x_0$; $t_0 \leftarrow 0$.
for $k \leftarrow 0, 1, \ldots$ **do**
  $J_k \leftarrow \max\big\{1, 2\log_2(\alpha_k D n / G)\big\}$; $N_k \leftarrow \big(4C(1 + \beta/\alpha_k)\big)^{1/\gamma}$; $t_{k+1} \leftarrow t_k + N_k J_k$.
  $x_{t_k}, x_{t_k+1}, \ldots, x_{t_{k+1}-1} \leftarrow y_k$.
  $y_{k,0} \leftarrow y_k$.
  for $j \leftarrow 0$ **to** $J_k - 1$ **do**
    $y_{k,j+1} \leftarrow \mathcal{A}(L_S(x) + \frac{\alpha_k}{2}\|x - x_0\|^2, \alpha_k + \beta, y_{k,j}, N_k)$.
  end
  $y_{k+1} \leftarrow y_{k,J_k}$; $\alpha_{k+1} \leftarrow \alpha_k / 2$.
end

Our general black-box conversion procedure, we name $\mathsf{USOL2}$, is presented in Algorithm 1. It is based on running the base algorithm $\mathcal{A}$ in epochs, where in each epoch we optimize a regularized version of the empirical risk, $L_S^{(k)}(x) = L_S(x) + \alpha_k \|x - x_0\|^2/2$. To do so, we invoke $\mathcal{A}$ for J_k times, each time halving the distance to the regularized minimizer by performing N_k steps. Between epochs, we halve the regularization magnitude.

The basic idea behind USOL2 is simple: in each epoch the algorithm converges close to the minimizer of the regularized objective $L_S^{(k)}(x)$, which is a stable function of the sample thanks to the added strongly convex regularization. The convergence to the minimizer can be exponentially fast since $L_S^{(k)}(x)$ is strongly convex and smooth, thus the output of each epoch is also stable. Following each epoch we decrease the regularization, thus converging closer and closer to the actual minimizer of the empirical (unregularized) risk while still maintaining uniform stability.

Note that if $\mathcal{A}$ is a first-order algorithm that uses only gradient access to the empirical risk (such as GD or NAG), then USOL2($\mathcal{A}$) can also be implemented using first-order access, as we only add a simple L2 regularization term to the empirical risk whose gradient is easy to compute.

The main result of this section is the following theorem which establishes convergence and stability guarantees for USOL2.

Theorem 6. *Assume $\ell(\cdot, z)$ is convex, β-smooth and G-Lipschitz w.r.t. $\|\cdot\|_2$ on $\mathcal{X}$. Let $\mathcal{A}$ be an optimization algorithm with convergence rate $C\beta\|x_0 - x^\star\|^2/t^\gamma$. Then the iterates $\{x_t\}_t$ produced by USOL2($\mathcal{A}$) initialized at x_0 such that $D \geq \|x_0 - x^\star\|$ satisfy the following, for all t:*

(i) x_t *is* $O(G^2 t^\gamma / C\beta n)$*-uniformly stable.*

(ii) $L_S(x_t) - L_S(x^\star) = \widetilde{O}(C\beta D^2/t^\gamma)$.

Following are lemmas needed to prove Theorem 6. In the lemmas we will refer to the values computed in USOL2 (such as y_k, x_t, etc.) for an arbitrary input. We will use the notation $L_S^{(k)}(x) = L_S(x) + \frac{\alpha_k}{2}\|x - x_0\|^2$ for the regularized objective we optimize at iteration k of the algorithm, and $x_k^\star$ to be its minimizer. Deferred proofs can be found in Appendix A.

We start with a technical lemma relating a regularized minimizer to the minimizer of the regularization and the minimizer of the function.

Lemma 7. *Let $f(x)$ be a convex function and let $f^{(\alpha)}(x) = f(x) + \frac{\alpha}{2}\|x - x_0\|^2$ for some $\alpha > 0$ and x_0. Let $x^\star \in \arg\min_x f(x)$ and $x_\alpha^\star \in \arg\min_x f^{(\alpha)}(x)$. Then*

$$\|x_0 - x_\alpha^\star\|^2 + \|x_\alpha^\star - x^\star\|^2 \leq \|x_0 - x^\star\|^2.$$

Proof. It is enough to show that $(x_0 - x_\alpha^\star) \cdot (x_\alpha^\star - x^\star) \geq 0$. To see this, note that $\nabla f^{(\alpha)}(x_\alpha^\star) = 0$ implies $\nabla f(x_\alpha^\star) + \alpha(x_\alpha^\star - x_0) = 0$, thus $x_0 - x_\alpha^\star = \frac{1}{\alpha}\nabla f(x_\alpha^\star)$. By convexity,

$$(x_0 - x_\alpha^\star) \cdot (x_\alpha^\star - x^\star) = \tfrac{1}{\alpha}\nabla f(x_\alpha^\star) \cdot (x_\alpha^\star - x^\star) \geq \tfrac{1}{\alpha}\big(f(x_\alpha^\star) - f(x^\star)\big) \geq 0. \qquad \blacksquare$$

Following lemma present the convergence in both function value and parameter distance of y_{k+1} with respect to $L_S^{(k)}$.

Lemma 8. *For all $k \geq 0$,*

$$L_S^{(k)}(y_{k+1}) - L_S^{(k)}(x_k^\star) \leq \frac{\alpha_k\|y_k - x_k^\star\|^2}{2^{J_k+1}} \qquad \text{and} \qquad \|y_{k+1} - x_k^\star\|^2 \leq \frac{\|y_k - x_k^\star\|^2}{2^{J_k}}.$$

Each time we update the regularization we start from the last y_k iteration. The following lemma bounds the distance of the last iteration to the new minimizer.

Lemma 9. *For all $k \geq 0$ it holds that $\|y_k - x_k^\star\| \leq \|x_0 - x_k^\star\|$.*

Next lemma yields the convergence guarantee of the y_k sequence.

Lemma 10. *For all $k \geq 0$,*

$$L_S(y_{k+1}) - L_S(x^\star) \leq \frac{3\alpha_k \|x_0 - x^\star\|^2}{4}.$$

Proof of Lemma 10. By the strong convexity of $L_S^{(k)}$,

$$\begin{aligned} L_S^{(k)}(y_{k+1}) - L_S^{(k)}(x^\star) &= L_S^{(k)}(y_{k+1}) - L_S^{(k)}(x_k^\star) + L_S^{(k)}(x_k^\star) - L_S^{(k)}(x^\star) \\ &\leq L_S^{(k)}(y_{k+1}) - L_S^{(k)}(x_k^\star) - \frac{\alpha_k}{2}\|x_k^\star - x^\star\|^2. \end{aligned}$$

Hence, using the definition of $L_S^{(k)}$,

$$\begin{aligned} L_S(y_{k+1}) - L_S(x^\star) &\leq L_S^{(k)}(y_{k+1}) - L_S^{(k)}(x^\star) + \frac{\alpha_k}{2}\|x^\star - x_0\|^2 \\ &\leq L_S^{(k)}(y_{k+1}) - L_S^{(k)}(x_k^\star) + \frac{\alpha_k}{2}\left(\|x^\star - x_0\|^2 - \|x_k^\star - x^\star\|^2\right) \\ &\leq \frac{\alpha_k \|y_k - x_k^\star\|}{4} + \frac{\alpha_k \|x^\star - x_0\|^2}{2}. \qquad \text{(Lemma 8 and } J_k \geq 1\text{)} \end{aligned}$$

By Lemma 9,

$$L_S(y_{k+1}) - L_S(x^\star) \leq \frac{\alpha_k}{4}\left(\|x_0 - x_k^\star\|^2 + 2\|x^\star - x_0\|^2\right).$$

We conclude by applying Lemma 7. ■

The following standard lemma bounds the distance between the minimizers of two functions where one of them is strongly convex.

Lemma 11. *Let $f_1, f_2 : \mathcal{X} \mapsto \mathbb{R}$ be convex and α-strongly convex functions (respectively) defined over a closed and convex domain $\mathcal{X} \subseteq \mathbb{R}^d$, and let $x_1 \in \arg\min_{x \in \mathcal{X}} f_1(x)$ and $x_2 \in \arg\min_{x \in \mathcal{X}} f_2(x)$. Then for $h = f_2 - f_1$ we have*

$$\|x_2 - x_1\| \leq \frac{2}{\alpha}\|\nabla h(x_1)\|_*.$$

Following is the stability guarantee for the y_k iterates of USOL2.

Lemma 12. *For $k \geq 0$, the iterate y_{k+1} produced by USOL2 is $(6G^2/n\alpha_k)$-uniformly stable.*

Proof of Lemma 12. Let $S = (z_1, \ldots, z_n)$ and $S' = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)$ be two neighboring datasets and let $\{y_k\}_k, \{\tilde{y}_k\}_k$ be the two "y_k" iterates obtained from USOL2 respectively. Using the triangle inequality,

$$\|y_{k+1} - \tilde{y}_{k+1}\| \leq \|y_{k+1} - x_k^\star\| + \|x_k^\star - \tilde{x}_k^\star\| + \|\tilde{x}_k^\star - \tilde{y}_{k+1}\|.$$

Using Lemma 8 and the definition of J_k,

$$\|y_{k+1} - x_k^\star\| \leq \frac{\|y_k - x_k^\star\|}{2^{J_k/2}} \leq \frac{G\|y_k - x_k^\star\|}{Dn\alpha_k}.$$

Combining Lemma 9 and Lemma 7,

$$\|y_k - x_k^\star\| \leq \|x_k^\star - x_0\| \leq \|x^\star - x_0\| \leq D.$$

Thus $\|y_{k+1} - x_k^\star\| \leq \frac{G}{n\alpha_k}$, and similarly $\|\tilde{x}_k^\star - \tilde{y}_{k+1}\| \leq \frac{G}{n\alpha_k}$. Hence,

$$\|y_{k+1} - \tilde{y}_{k+1}\| \leq \|x_k^\star - \tilde{x}_k^\star\| + \frac{2G}{n\alpha_k}.$$

We will now focus on bounding $\|x_k^\star - \tilde{x}_k^\star\|$. By invoking Lemma 11 with $L_S^{(k)}$ and $L_{S'}^{(k)}$,

$$\|x_k^\star - \tilde{x}_k^\star\| \leq \frac{2\|\ell(x_k^\star; z_i') - \ell(x_k^\star; z_i)\|_*}{n\alpha_k} \leq \frac{4G}{n\alpha_k},$$

where we have used the fact that $\ell(\cdot, z_i)$ and $\ell(\cdot, z_i')$ are G-Lipschitz. Thus, $\|y_{k+1} - \tilde{y}_{k+1}\| \leq \frac{6G}{n\alpha_k}$. Again using the fact that $\ell(\cdot, z)$ is Lipschitz,

$$\sup_{z\in\mathcal{Z}} |\ell(y_{k+1}; z) - \ell(\tilde{y}_{k+1}; z)| \leq G\|y_{k+1} - \tilde{y}_{k+1}\| \leq \frac{6G^2}{n\alpha_k},$$

hence we establish uniform stability. ■

We are now ready to prove Theorem 6.

Proof of Theorem 6. If $0 < t < t_1$ then $y_t = x_0$ and uniform stability is immediate. Regarding convergence, by smoothness,

$$L_S(x_0) - L_S(x^\star) \leq \frac{\beta D^2}{2} \leq \frac{\beta D^2 (N_0 J_0)^\gamma}{2t^\gamma} = \widetilde{O}\left(\frac{C\beta D^2}{t^\gamma}\right).$$

Let $K = \max\{k : t_{k+1} \leq t\}$ for some $t > 0$ which implies $x_t = y_{K+1}$. To obtain uniform stability by Lemma 12 we need to lower bound α_K. Thus,

$$t \geq t_{K+1} \geq N_K J_K \geq \left(\frac{4C\beta}{\alpha_K}\right)^{1/\gamma} \implies \alpha_k = \Omega\left(\frac{C\beta}{t^\gamma}\right).$$

Thus, x_t is $O\big(G^2 t^\gamma / C\beta n\big)$-uniformly stable. For the convergence result we need to upper bound α_K and invoke Lemma 10. First we bound N_k by a geometric series,

$$N_k = \left(4C\left(\frac{\beta}{\alpha_k} + 1\right)\right)^{1/\gamma} = \left(4C\left(2^{k+2} + 1\right)\right)^{1/\gamma} \leq \left(20C \cdot 2^k\right)^{1/\gamma} \implies \sum_{i=0}^{k} N_i \leq (20C)^{1/\gamma} \frac{2^{(k+1)/\gamma}}{2^{1/\gamma} - 1}.$$

Thus, using the definition of J_k,

$$t < t_{K+2} = \sum_{k=0}^{K+1} J_k N_k \leq \max\left\{1, 2\log_2 \frac{\beta D n}{4G}\right\} \sum_{k=0}^{K+1} N_k \leq (20C)^{1/\gamma} \max\left\{1, 2\log_2 \frac{\beta D n}{4G}\right\} \frac{2^{(K+2)/\gamma}}{2^{1/\gamma} - 1}.$$

Rearranging the terms and using $0 < \gamma \leq 2$,

$$\frac{1}{2^{K+2}} \leq \frac{10C}{t^\gamma \left(1 - 2^{-1/\gamma}\right)^\gamma} \max\left\{1, 2\log_2^\gamma \frac{\beta D n}{4G}\right\} = O\left(\frac{C}{t^\gamma} \max\left\{1, 2\log_2^\gamma \frac{\beta D n}{4G}\right\}\right).$$

We conclude using Lemma 10 with $x_t = y_{K+1}$ and $\alpha_K = \beta / 2^{K+2}$,

$$L_S(x_t) - L_S(x^\star) \leq \frac{3\alpha_K D^2}{4} = O\left(\frac{C\beta D^2}{t^\gamma} \max\left\{1, \log^\gamma\left(\frac{\beta D n}{G}\right)\right\}\right).$$ ■

4. Stable Mirror Descent for general norms

In this section we provide a uniformly stable variant of Mirror Descent for empirical risk minimization in general normed spaces. The algorithm, which we term USMD is presented in Algorithm 2). Here we assume that the loss function $\ell : \mathcal{X} \times \mathcal{Z} \mapsto \mathbb{R}$ is convex, β-smooth and G-Lipschitz (in its first argument) w.r.t. a general $\|\cdot\|$.

Like the standard Mirror Descent, the algorithm is parameterized by a regularization function $R(x)$ which is 1-strongly convex w.r.t. $\|\cdot\|$. Let $x_0 = \arg\min_{x \in \mathcal{X}} R(x)$, $x^\star \in \arg\min_{x \in \mathcal{X}} L_S(x)$ and let $D^2 \geq R(x^\star) - R(x_0)$. In each step of USMD we perform linearization of $L_S(x_t)$ and use $\alpha R(x)$

Algorithm 2: USMD: Uniformly Stable Mirror Descent

Input: Sample $S = (z_i)_{i=1}^n$, smoothness β, x_0, regularization parameter α, number of steps T.
for $t \leftarrow 0$ **to** $T-1$ **do**

$$x_{t+1} \leftarrow \arg\min_{x \in \mathcal{X}} \{\nabla L_S(x_t) \cdot (x - x_t) + \beta D_R(x, x_t) + \alpha R(x)\}.$$

end
Output: x_T.

as a regularization in addition to the Bregman of the mirror descent step, which lets us obtain linear rate convergence on the regularized objective. Hence, we obtain stability by converging near to a regularized minimizer at a minimal cost. α is carefully tuned to balance the stability and empirical error. Note that the update step of USMD is in fact a Mirror Descent step on the regularized function $L_S^{(\alpha)}(x) \triangleq L_S(x) + \alpha R(x)$ with smoothness of $\alpha + \beta$. This can be seen by comparing the arg min step of USMD and that of Mirror Descent (Eq. (1)), and is written formally in the proof of Theorem 13.

Further, note that first-order access to the empirical risk suffices for implementing our method as we only access it by performing linearization in each step. Following is the main result for this section, describing the stability and convergence of Algorithm 2.

Theorem 13. *Assume $\ell(\cdot, z)$ is convex, β-smooth and G-Lipschitz w.r.t. $\|\cdot\|$ on $\mathcal{X}$, $R(x)$ is 1-strongly convex w.r.t. $\|\cdot\|$ on $\mathcal{X}$, $x_0 = \arg\min_{x \in \mathcal{X}} R(x)$ and $D^2 \geq R(x^\star) - R(x_0)$. Then given $T \geq 2 \log \frac{\beta D n}{G}$, the output of USMD (the final iterate x_T) with $\alpha = \frac{\beta}{T} \max\left\{1, 2 \log_2 \frac{\beta D n}{GT}\right\}$ satisfies the following:*

- *x_T is $O(G^2 T / \beta n)$-uniformly stable.*
- *$L_S(x_T) - L_S(x^\star) = \widetilde{O}(\beta D^2 / T)$.*

The following lemmas are used in order to prove Theorem 13. We defer their proofs to Appendix B. As we mentioned, adding regularization can impair smoothness (cf. Appendix C), hence we cannot appeal directly to classical bounds for smooth Mirror Descent. The next lemma show that adding regularization, although not necessarily smooth, is indeed relatively smooth and strongly convex.

Lemma 14. *Let $f(x)$ be a convex and β-smooth function w.r.t. $\|\cdot\|$. Let $R(x)$ be 1-strongly convex w.r.t. $\|\cdot\|$. Then $f^{(\alpha)}(x) \triangleq f(x) + \alpha R(x)$ for $\alpha > 0$ is $(\alpha + \beta)$-smooth and α-strongly convex relative to $R(x)$.*

Hence, one can take advantage of the method of Lu et al. (2018) for relatively-smooth convex optimization. The following lemma is derived from the analysis of Lu et al. (2018).

Lemma 15. *Let $f(x)$ be β-smooth and α-strongly convex relative to $R(x)$. Then the sequence $\{x_t\}$ defined by Eq. (1) satisfy:*

(i) $\{f(x_t)\}_t$ *is monotonically decreasing.*

(ii) Let $x^\star \in \arg\min_{x\in\mathcal{X}} f(x)$*. For all* $t \geq 1$, $D_R(x^\star, x_t) \leq \left(1 - \frac{\alpha}{\beta}\right)^t D_R(x^\star, x_0)$.

(iii) For all $t \geq 1$ *and* $x \in \mathcal{X}$, $f(x_t) - f(x) \leq \frac{\alpha D_R(x, x_0)}{(1+\frac{\alpha}{\beta-\alpha})^t - 1}$.

We are now ready to prove Theorem 13.

Proof of Theorem 13. We start with showing that $D_R(x^\star_\alpha, 0) \leq D^2$. This inequality will be used for both stability and convergence results. By the zero-order optimality of $x^\star_\alpha$ and $x^\star$,

$$
\begin{aligned}
L_S(x^\star_\alpha) + \alpha R(x^\star_\alpha) &\leq L_S(x^\star) + \alpha R(x^\star) \\
\implies R(x^\star) - R(x^\star_\alpha) &\geq \frac{L_S(x^\star_\alpha) - L_S(x^\star)}{\alpha} \geq 0.
\end{aligned}
$$

From the first-order optimality of x_0 which implies $\nabla R(x_0) \cdot (x_0 - x^\star_\alpha) \leq 0$,

$$D_R(x^\star_\alpha, x_0) = R(x^\star_\alpha) - R(x_0) - \nabla R(x_0) \cdot (x^\star_\alpha - x_0) \leq R(x^\star_\alpha) - R(x_0).$$

Combining the two inequalities,

$$D_R(x^\star_\alpha, x_0) \leq R(x^\star_\alpha) - R(x_0) \leq R(x^\star) - R(x_0) \leq D^2. \tag{2}$$

Secondly we will show that our method in fact performs mirror steps on $L_S^{(\alpha)}(x) = L_S(x) + \alpha R(x)$ which is $(\alpha+\beta)$-smooth and α-strongly convex relative to $R(x)$ by Lemma 14. The update step of USMD is

$$x_{t+1} = \arg\min_{x\in\mathcal{X}} \{\nabla L_S(x_t) \cdot (x - x_t) + \beta D_R(x, x_t) + \alpha R(x)\}.$$

Using the definition of $D_R(x, x_t)$,

$$
\begin{aligned}
&\nabla L_S(x_t) \cdot (x - x_t) + \beta D_R(x, x_t) + \alpha R(x) \\
&= \nabla L_S(x_t) \cdot (x - x_t) + (\alpha+\beta) D_R(x, x_t) + \alpha R(x) - \alpha(R(x) - R(x_t) - \nabla R(x_t) \cdot (x - x_t)) \\
&= (\nabla L_S(x_t) + \alpha \nabla R(x_t)) \cdot (x - x_t) + (\alpha+\beta) D_R(x, x_t) + \alpha R(x_t).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\arg\min_{x\in\mathcal{X}} \left\{\nabla L_S(x_t) \cdot (x - x_t) + \beta D_R(x, x_t) + \alpha R(x)\right\} \\
&= \arg\min_{x\in\mathcal{X}} \left\{\nabla L_S^{(\alpha)}(x_t) \cdot (x - x_t) + (\alpha+\beta) D_R(x, x_t)\right\},
\end{aligned}
$$

which is a mirror descent step (Eq. (1)) for $L_S^{(\alpha)}$ with a smoothness of $\beta + \alpha$. Hence, we can invoke Lemma 15.

Next follows the stability argument. Let $S = (z_1, \dots, z_n)$ and $S' = (z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_n)$. Let x_T and $\tilde{x}_T$ be the outputs of USMD on S and S' respectively. Let $x_\alpha^\star = \arg\min_{x\in\mathcal{X}} L_S^{(\alpha)}(x)$ and $\tilde{x}_\alpha^\star = \arg\min_{x\in\mathcal{X}} L_{S'}^{(\alpha)}(x)$. Using the triangle inequality,

$$\|x_T - \tilde{x}_T\| \le \|x_T - x_\alpha^\star\| + \|x_\alpha^\star - \tilde{x}_\alpha^\star\| + \|\tilde{x}_\alpha^\star - \tilde{x}_T\|.$$

By Lemma 15,

$$D_R(x_\alpha^\star, x_T) \le \left(1 - \frac{\alpha}{\alpha+\beta}\right)^T D_R(x_\alpha^\star, x_0) = \left(\frac{1}{1+\frac{\alpha}{\beta}}\right)^T D_R(x_\alpha^\star, x_0).$$

Using the inequality $\left(1+\frac{1}{x}\right)^x \ge 2$ for $x \ge 1$, with $\frac{\beta}{\alpha} \ge 1$, $\left(1+\frac{\alpha}{\beta}\right)^{\beta/\alpha} \ge 2$. Note that $\frac{\beta}{\alpha} \ge 1$ due to our assumption that $2\log_2\frac{\beta D n}{G} \le T$. Thus,

$$D_R(x_\alpha^\star, x_T) \le \left(\frac{1}{1+\frac{\alpha}{\beta}}\right)^T D_R(x_\alpha^\star, x_0) \le 2^{-\frac{\alpha T}{\beta}} D_R(x_\alpha^\star, x_0).$$

By $\alpha \ge \frac{2\beta}{T}\log_2\frac{\beta D n}{GT}$ and Eq. (2),

$$D_R(x_\alpha^\star, x_T) \le \frac{G^2T^2}{\beta^2 D^2 n^2} D_R(x_\alpha^\star, x_0) \le \frac{G^2T^2}{\beta^2 n^2}.$$

From the strong convexity of $R(x)$, $\frac{1}{2}\|x_\alpha^\star - x_T\|^2 \le D_R(x_\alpha^\star, x_T)$. Hence,

$$\|x_\alpha^\star - x_T\| \le \frac{\sqrt{2}GT}{\beta n},$$

and similarly $\|\tilde{x}_T - \tilde{x}_\alpha^\star\| \le \frac{\sqrt{2}GT}{\beta n}$. Now we will bound $\|x_\alpha^\star - \tilde{x}_\alpha^\star\|$. Using Lemma 11 with $f_1 = L_S^{(\alpha)}$ (α-strongly convex since $R(x)$ is 1-strongly convex) and $f_2 = L_{S'}^{(\alpha)}$,

$$\|x_\alpha^\star - \tilde{x}_\alpha^\star\| \le \frac{2\|\ell(x_\alpha^\star; z_i') - \ell(x_\alpha^\star; z_i)\|_*}{n\alpha} \le \frac{4G}{n\alpha},$$

where we have used the fact that $\ell(\cdot, z_i)$ and $\ell(\cdot, z_i')$ are G-Lipschitz. Thus, since $\alpha \ge \beta/T$,

$$\|x_T - \tilde{x}_T\| \le \frac{4G}{n\alpha} + \frac{2\sqrt{2}GT}{n\beta} \le \frac{(4+2\sqrt{2})GT}{n\beta}.$$

Since $\ell(\cdot, z)$ is G-Lipschitz, we upper bound the uniform stability,

$$\sup_{z\in\mathcal{Z}}|\ell(x_T; z) - \ell(\tilde{x}_T; z)| \le G\|x_T - \tilde{x}_T\| \le \frac{(4+2\sqrt{2})G^2T}{n\beta} = O\left(\frac{G^2T}{\beta n}\right).$$

We move on to the convergence of x_T.

$$\begin{aligned} L_S(x_T) - L_S(x^\star) &= L_S^{(\alpha)}(x_T) - L_S^{(\alpha)}(x^\star) + \alpha(R(x^\star) - R(x_T)) && (L_S^{(\alpha)}(x) = L_S(x) + \alpha R(x)) \\ &\le L_S^{(\alpha)}(x_T) - L_S^{(\alpha)}(x_\alpha^\star) + \alpha(R(x^\star) - R(x_T)). && (x_\alpha^\star = \arg\min_{x\in\mathcal{X}} L_S^{(\alpha)}(x)) \end{aligned}$$

Again by Lemma 15, and $T \geq \frac{\beta}{\alpha}$ which implies $\left(1 + \frac{\alpha}{\beta}\right)^T \geq 2$,

$$L_S^{(\alpha)}(x_T) - L_S^{(\alpha)}(x_\alpha^\star) \leq \frac{\alpha D_R(x_\alpha^\star, x_0)}{\left(1 + \frac{\alpha}{\beta}\right)^T - 1} \leq \alpha D_R(x_\alpha^\star, x_0).$$

Thus, using the minimality of x_0,

$$L_S(x_T) - L_S(x^\star) \leq \alpha(D_R(x_\alpha^\star, x_0) + R(x^\star) - R(x_T)) \leq \alpha(D_R(x_\alpha^\star, x_0) + R(x^\star) - R(x_0)),$$

and by Eq. (2) and the definition of α,

$$L_S(x_T) - L_S(x^\star) \leq 2\alpha D^2 = \widetilde{O}\left(\frac{\beta D^2}{T}\right). \qquad \blacksquare$$

Acknowledgments

This work has received support from the Israeli Science Foundation (ISF) grant no. 2549/19, the Len Blavatnik and the Blavatnik Family foundation, the Deutsch Foundation, and the Yandex Initiative in Machine Learning.

References

Naman Agarwal, Rohan Anil, Tomer Koren, Kunal Talwar, and Cyril Zhang. Stochastic optimization with laggard data pipelines. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. In *International Conference on Machine Learning*, pages 393–403. PMLR, 2021.

Amit Attia and Tomer Koren. Algorithmic instabilities of accelerated gradient descent. *Advances in Neural Information Processing Systems*, 34, 2021.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 2019.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33, 2020.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 745–754. PMLR, 2018.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. *Advances in Neural Information Processing Systems*, 29:3576–3584, 2016.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, pages 9747–9757, 2018.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279, 2019.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

Alon Gonen and Shai Shalev-Shwartz. Fast rates for empirical risk minimization of strict saddle problems. In *Conference on Learning Theory*, pages 1043–1063. PMLR, 2017.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $O(1/n)$. *Advances in Neural Information Processing Systems*, 34, 2021.

Tomer Koren and Kfir Y Levy. Fast rates for exp-concave empirical risk minimization. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1477–1485, 2015.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

Ben London. A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Andreas Maurer. A second-order look at stability and generalization. In *Conference on learning theory*, pages 1461–1475. PMLR, 2017.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

Jialei Wang, Weiran Wang, and Nathan Srebro. Memory and communication efficient distributed stochastic optimization with minibatch prox. In *Conference on Learning Theory*, pages 1882–1919. PMLR, 2017.

Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1307–1322, 2017.

Appendix A. Proofs of Section 3

A.1. Proof of Lemma 8

Proof. From strong convexity and the rate of $\mathcal{A}$,

$$\frac{\alpha_k}{2}\|y_{k,j+1} - x_k^\star\|^2 \le L_S^{(k)}(y_{k,j+1}) - L_S^{(k)}(x_k^\star) \le \frac{C(\alpha_k+\beta)\|y_{k,j} - x_k^\star\|^2}{4C\left(1+\frac{\beta}{\alpha_k}\right)} = \frac{\alpha_k\|y_{k,j} - x_k^\star\|^2}{4}.$$

Hence,

$$\|y_{k,j+1} - x_k^\star\|^2 \le \frac{\|y_{k,j} - x_k^\star\|^2}{2}.$$

Repeating this argument and substituting $y_k = y_{k,0}$,

$$\|y_{k,j+1} - x_k^\star\|^2 \le \frac{\|y_{k,0} - x_k^\star\|^2}{2^{j+1}} = \frac{\|y_k - x_k^\star\|^2}{2^{j+1}}.$$

Thus,

$$L_S^{(k)}(y_{k,J_k}) - L_S^{(k)}(x_k^\star) \le \frac{C(\alpha_k + \beta)\|y_{k,J_k-1} - x_k^\star\|^2}{4C(1+\beta/\alpha_k)} = \frac{\alpha_k\|y_{k,J_k-1} - x_k^\star\|^2}{4} \le \frac{\alpha_k\|y_k - x_k^\star\|^2}{2^{J_k+1}}.$$

We conclude by substituting $y_{k+1} = y_{k,J_k}$. ■

A.2. Proof of Lemma 9

First we need the following claim, proved below.

Claim 16. *Let $f(x)$ a convex function. Let $x_1^\star$ and $x_2^\star$ be the minimizers of $f(x) + \frac{\alpha_1}{2}\|x - x_0\|^2$ and $f(x) + \frac{\alpha_2}{2}\|x - x_0\|^2$ respectively, for some $x_0 \in \mathbb{R}^d$ and $\alpha_1, \alpha_2 > 0$. Then,*

$$\|x_1^\star - x_2^\star\|^2 \le \frac{\alpha_1 - \alpha_2}{\alpha_1 + \alpha_2}\Big(\|x_2^\star - x_0\|^2 - \|x_1^\star - x_0\|^2\Big).$$

Proof of Lemma 9. For $k = 0$ the claim is immediate as $y_0 = x_0$. For $k > 0$, using the triangle inequality,

$$\|y_k - x_k^\star\| \le \|y_k - x_{k-1}^\star\| + \|x_{k-1}^\star - x_k^\star\|.$$

By Claim 16,

$$\|x_{k-1}^\star - x_k^\star\|^2 \le \frac{\alpha_{k-1} - \alpha_k}{\alpha_{k-1} + \alpha_k}\Big(\|x_k^\star - x_0\|^2 - \|x_{k-1}^\star - x_0\|^2\Big) \le \frac{\|x_k^\star - x_0\|^2 - \|x_{k-1}^\star - x_0\|^2}{3}.$$

Due to Lemma 8,

$$\begin{aligned}
\|y_k - x_{k-1}^\star\| &\le \frac{\|y_{k-1} - x_{k-1}^\star\|}{2^{J_k/2}} \\
&\le \frac{\|x_0 - x_{k-1}^\star\|}{2^{J_k/2}} && \text{(induction)} \\
&\le \frac{\|x_0 - x_{k-1}^\star\|}{\sqrt{2}}. && (J_k \ge 1)
\end{aligned}$$

Thus,

$$\begin{aligned}
\|y_k - x_k^\star\| &\le \frac{\|x_0 - x_{k-1}^\star\|}{\sqrt{2}} + \sqrt{\frac{\|x_k^\star - x_0\|^2 - \|x_{k-1}^\star - x_0\|^2}{3}} \\
&= \|x_k^\star - x_0\|\left(\frac{\|x_0 - x_{k-1}^\star\|/\|x_k^\star - x_0\|}{\sqrt{2}} + \sqrt{\frac{1 - \|x_{k-1}^\star - x_0\|^2/\|x_k^\star - x_0\|^2}{3}}\right).
\end{aligned}$$

From Claim 16, $\|x_{k-1}^\star - x_0\| \le \|x_k^\star - x_0\|$, and since for $0 \le x \le 1$, $\frac{x}{\sqrt{2}} + \sqrt{\frac{1-x^2}{3}} \le 1$,

$$\|y_k - x_k^\star\| \le \|x_k^\star - x_0\|.$$ ■

Proof of Claim 16. From the strong convexity of $f(x) + \frac{\alpha_1}{2}\|x - x_0\|^2$,

$$\frac{\alpha_1}{2}\|x_2^\star - x_1^\star\|^2 \le f(x_2^\star) + \frac{\alpha_1}{2}\|x_2^\star - x_0\|^2 - f(x_1^\star) - \frac{\alpha_1}{2}\|x_1^\star - x_0\|^2.$$

Similarly,

$$\frac{\alpha_2}{2}\|x_1^\star - x_2^\star\|^2 \le f(x_1^\star) + \frac{\alpha_2}{2}\|x_1^\star - x_0\|^2 - f(x_2^\star) - \frac{\alpha_2}{2}\|x_2^\star - x_0\|^2$$
$$\implies f(x_2^\star) - f(x_1^\star) \le \frac{\alpha_2}{2}\|x_1^\star - x_0\|^2 - \frac{\alpha_2}{2}\|x_2^\star - x_0\|^2 - \frac{\alpha_2}{2}\|x_1^\star - x_2^\star\|^2.$$

Combining the two inequalities,

$$\frac{\alpha_1}{2}\|x_2^\star - x_1^\star\|^2 \le \frac{\alpha_1 - \alpha_2}{2}\left(\|x_2^\star - x_0\|^2 - \|x_1^\star - x_0\|^2\right) - \frac{\alpha_2}{2}\|x_1^\star - x_2^\star\|^2,$$

and we obtain the desired result by rearranging the terms. ■

A.3. Proof of Lemma 11

Proof. The strong convexity of f_2 and x_2 being the minimum of f_2 implies

$$\nabla f_2(x_1) \cdot (x_1 - x_2) \ge f_2(x_1) - f_2(x_2) + \frac{\alpha}{2}\|x_2 - x_1\|^2 \ge \frac{\alpha}{2}\|x_2 - x_1\|^2.$$

From first-order optimality of x_1, $\nabla f_1(x_1) \cdot (x_1 - x_2) \le 0$. Thus,

$$\nabla f_2(x_1) \cdot (x_1 - x_2) = \nabla f_1(x_1) \cdot (x_1 - x_2) + \nabla h(x_1) \cdot (x_1 - x_2) \le \nabla h(x_1) \cdot (x_1 - x_2).$$

Putting the two inequalities, together with Hölder inequality, yields

$$\frac{\alpha}{2}\|x_2 - x_1\|^2 \le \nabla h(x_1) \cdot (x_1 - x_2) \le \|\nabla h(x_1)\|_* \|x_1 - x_2\|.$$

Thus, $\|x_2 - x_1\| \le \frac{2}{\alpha}\|\nabla h(x_1)\|_*$. ■

Appendix B. Proofs of Section 4

B.1. Proof of Lemma 14

Proof. Due to convexity of f, $f(y) \ge f(x) + \nabla f(x) \cdot (y - x)$. Since $D_R(y, x) = R(y) - R(x) - \nabla R(x) \cdot (y - x)$,

$$f(y) + \alpha R(y) \ge f(x) + (\nabla f(x) + \alpha \nabla R(x)) \cdot (y - x) + \alpha R(x) + \alpha D_R(y, x).$$

Hence, by the definition of $f^{(\alpha)}(x)$,

$$f^{(\alpha)}(y) \ge f^{(\alpha)}(x) + \nabla f^{(\alpha)}(x) \cdot (y - x) + \alpha D_R(y, x),$$

and we conclude that $f^{(\alpha)}(x)$ is α-strongly convex relative to $R(x)$. Since $f(x)$ is β-smooth,

$$f(y) \le f(x) + \nabla f(x) \cdot (y - x) + \frac{\beta}{2}\|y - x\|^2.$$

Using the inequality $D_R(y,x) \geq \frac{1}{2}\|y-x\|^2$ (since R is 1-strongly convex),

$$f(y) \leq f(x) + \nabla f(x) \cdot (y-x) + \beta D_R(y,x).$$

Adding $\alpha D_R(y,x)$ to both sides and using the definition of $D_R(y,x)$,

$$f(y) + \alpha R(y) - \alpha R(x) - \alpha \nabla R(x) \cdot (y-x) \leq f(x) + \nabla f(x) \cdot (y-x) + (\alpha+\beta) D_R(y,x).$$

Hence, by the definition of $f^{(\alpha)}(x)$,

$$f^{(\alpha)}(y) \leq f^{(\alpha)}(x) + \nabla f^{(\alpha)}(x) \cdot (y-x) + (\alpha+\beta) D_R(y,x),$$

and we conclude that $f^{(\alpha)}(x)$ is $(\alpha+\beta)$-smooth relative to $R(x)$. ■

B.2. Proof of Lemma 15

The lemma is based on the analysis of Lu et al. (2018) (Theorem 3.1) which gives the convergence in terms of $f(x_t) - f(x)$. For Theorem 13 we also need convergence in terms of $\frac{D_R(x^\star, x_t)}{D_R(x^\star, x_0)}$ for $x^\star \in \arg\min_{x \in \mathcal{X}} f(x)$. Thus, we repeat the argument for completeness. The proof relies on the following Three-Point Property:

Lemma 17 (Three-Point Property of Tseng (2008)). *Let $\phi(x)$ be a convex function, and let $D_R(\cdot,\cdot)$ be the Bregman distance for $R(\cdot)$. For a given vector z, let*

$$z^+ \triangleq \arg\min_{x \in \mathcal{X}} \{\phi(x) + D_R(x,z)\}.$$

Then

$$\phi(x) + D_R(x,z) \geq \phi(z^+) + D_R(z^+,z) + D_R(x,z^+) \text{ for all } x \in \mathcal{X}.$$

Proof of Lemma 15. For any $x \in \mathcal{X}$ and $t \geq 1$, from relative smoothness,

$$f(x_t) \leq f(x_{t-1}) + \nabla f(x_{t-1}) \cdot (x_t - x_{t-1}) + \beta D_R(x_t, x_{t-1}).$$

Using Lemma 17 with $\phi(x) = \frac{1}{\beta} \nabla f(x_{t-1}) \cdot (x - x_{t-1})$ and $z = x_{t-1}$,

$$\frac{1}{\beta} \nabla f(x_{t-1}) \cdot (x - x_t) + D_R(x, x_{t-1}) \geq D_R(x_t, x_{t-1}) + D_R(x, x_t).$$

Hence,

$$f(x_t) \leq f(x_{t-1}) + \nabla f(x_{t-1}) \cdot (x - x_{t-1}) + \beta D_R(x, x_{t-1}) - \beta D_R(x, x_t).$$

Thus, from relative strong convexity,

$$f(x_t) \leq f(x) + (\beta - \alpha) D_R(x, x_{t-1}) - \beta D_R(x, x_t).$$

Note that if $x = x_{t-1}$, $f(x_t) \leq f(x_{t-1})$. Hence, $\{f(x_t)\}_t$ is monotonically decreasing. From the definition of $x^\star$, $f(x_t) \geq f(x^\star)$. Hence, for $x = x^\star$ we obtain

$$D_R(x^\star, x_t) \leq \frac{\beta - \alpha}{\beta} D_R(x^\star, x_{t-1}) = \left(1 - \frac{\alpha}{\beta}\right) D_R(x^\star, x_{t-1}).$$

Repeating this argument and we obtain

$$D_R(x^\star, x_t) \le \left(1 - \frac{\alpha}{\beta}\right)^t D_R(x^\star, x_0).$$

Returning to

$$f(x_t) \le f(x) + (\beta - \alpha) D_R(x, x_{t-1}) - \beta D_R(x, x_t),$$

it follows by induction that

$$\sum_{i=1}^{t} \left(\frac{\beta}{\beta - \alpha}\right)^i (f(x_i) - f(x)) \le \beta D_R(x, x_0) - \left(\frac{\beta}{\beta - \alpha}\right)^t \beta D_R(x, x_t).$$

Since $\{f(x_t)\}_t$ is monotonically decreasing,

$$\begin{aligned}\sum_{i=1}^{t} \left(\frac{\beta}{\beta - \alpha}\right)^i (f(x_i) - f(x)) &\ge \sum_{i=1}^{t} \left(\frac{\beta}{\beta - \alpha}\right)^i (f(x_t) - f(x)) \\ &= \frac{\beta\left(1 + \frac{\alpha}{\beta - \alpha}\right)^t - 1}{\alpha} (f(x_t) - f(x)).\end{aligned}$$

Thus,

$$\begin{aligned}\frac{\beta\left(1 + \frac{\alpha}{\beta - \alpha}\right)^t - 1}{\alpha} (f(x_t) - f(x)) &\le \beta D_R(x, x_0) - \left(\frac{\beta}{\beta - \alpha}\right)^t \beta D_R(x, x_t) \\ &\le \beta D_R(x, x_0).\end{aligned}$$

Rearranging the terms yields the convergence result. ■

Appendix C. On Smoothness of ℓ_p Regularization

The following lemma indicate that $\|\cdot\|_p^2$ for $1 < p < 2$ is not smooth. Hence, using it as a regularization can impair a smoothness assumption.

Lemma 18. *Let $f(x) = \|x\|_p^2$ for $1 < p < 2$ over $\mathcal{X} = \mathbb{R}^d$ ($d > 1$). Then $f(x)$ is not smooth with respect to $\|\cdot\|_p$.*

Proof. Assume by contradiction that $R(x)$ is β-smooth with respect to $\|\cdot\|_p$ for some $\beta > 0$. Let $x = (1, 0, 0, \dots, 0)^T$ and $y = (1, \epsilon, 0, \dots, 0)^T$ for some $\epsilon > 0$. Thus,

$$\begin{aligned}f(y) - f(x) - \nabla f(x) \cdot (y - x) &= (1 + \epsilon^p)^{2/p} - 1 - \nabla f(x) \cdot (y - x) \\ &= (1 + \epsilon^p)^{2/p} - 1 - (1, 0, \dots, 0)^T \cdot (0, \epsilon, 0, \dots, 0)^T \\ &= (1 + \epsilon^p)^{2/p} - 1.\end{aligned}$$

Using the identity $(1 + x)^\gamma \ge 1 + \gamma x$ for $x \ge 0$ and $\gamma \ge 1$, $(1 + \epsilon^p)^{2/p} - 1 \ge \frac{2}{p}\epsilon^p$. Since $p < 2$,

$$\lim_{\epsilon \to 0^+} \frac{\epsilon^2}{\epsilon^p} = 0.$$

We can pick sufficiently small ϵ with

$$\frac{\epsilon^2}{(1+\epsilon^p)^{2/p}-1} \le \frac{p\epsilon^2}{2\epsilon^p} < \frac{2}{\beta},$$

for which

$$f(y) - f(x) - \nabla f(x) \cdot (y-x) = (1+\epsilon^p)^{2/p} - 1 > \frac{\beta\epsilon^2}{2} = \frac{\beta}{2}\|y-x\|_p^2,$$

and we get a contradiction to the smoothness assumption. ■