# CPSC 532D — 13. STABILITY, REGULARIZATION, AND CONVEX PROBLEMS

*Danica J. Sutherland*

*University of British Columbia, Vancouver*

*Fall 2023*

---

We've argued that it's important to consider algorithm-specific bounds: sometimes $\mathcal{H}$ is too big for uniform convergence, but $\mathcal{A}$ still learns well. We motivated this last time from deep learning, but this is true even for, say, kernel ridge regression with a universal kernel,

$$\arg\min_{h \in \mathcal{F}} \mathrm{L}_\mathrm{S}^{sq}(h) + \lambda \|h\|_{\mathcal{F}}^2 .$$

This algorithm could feasibly return any $h \in \mathcal{F}$, and so a typical uniform convergence analysis will have to use $\mathcal{H} = \mathcal{F}$. But we know that the VC dimension and Rademacher complexity of $\mathcal{F}$ are both infinite for a universal kernel, so a uniform convergence analysis won't tell us anything useful.

*Consider a $\mathcal{D}$ where $y \mid x = h^*(x)$, the distribution on $x$ is "sufficiently broad," and $\lambda \to 0$ as $m \to \infty$. If this kernel is also continuous, $h^*$ is the unique minimizer of $\mathrm{L}_\mathcal{D}$, and checking the details in some asymptotic analysis will give us that $\hat{h}_\mathrm{S} \to h^*$.*

We can generalize this setting a bit to *regularized loss minimization* (RLM):

$$\arg\min_{h \in \mathcal{H}} \mathrm{L}_\mathrm{S}(h) + \lambda \mathrm{R}(h). \tag{RLM}$$

(Kernel ridge regression uses square loss, with $\mathcal{H}$ the RKHS $\mathcal{F}$, and $\mathrm{R}(h) = \|h\|_{\mathcal{F}}^2$.)

We've appealed before to a connection between RLM and constrained ERM:

$$\arg\min_{h \in \mathcal{H} : \mathrm{R}(h) \le \mathrm{B}} \mathrm{L}_\mathrm{S}(h).$$

The two problems are Lagrange dual to each other: for any $\lambda$, there is some B such that the solutions agree, and vice versa. But converting between $\lambda$ and B generally requires actually solving the problem, (RLM) is typically easier computationally, and it's usually easier to choose a good $\lambda$ than to choose a good B. So, we'd like to come up with a direct analysis of (RLM) for its own sake, in addition to developing techniques that will help with other analyses.

Randomized algorithms Our analyses of ERM applied to *any* empirical risk minimizer, so exactly what the algorithm did didn't really matter. When we're dealing with specific algorithms, though, it's important to note that these algorithms might be *randomized*: for instance, stochastic gradient descent sees points in a random order, and might start at a random location.

Notationally, we'll use $\mathcal{A}(\mathrm{S})$ to denote the *potentially random* result of the algorithm on a set S: even for a fixed S, the result could be a random variable. Something like $\mathbb{E}_\mathcal{A} \ell(\mathcal{A}(\mathrm{S}), z)$ will denote the mean of the loss on a fixed $z$ when training on a fixed S. We'll always assume that this randomness is independent of S (other than its size): for instance, this will be the random seed that determines that pattern in which we access the $z_i$. If $\mathcal{A}$ is deterministic, it's just a point-mass distribution.

For more, visit https://cs.ubc.ca/~dsuth/532D/23w1/.

## 1 Definitions of Stability

The general intuition we're going to use is *algorithmic stability*: if two training sets S and S′ are "similar," then $\mathcal{A}(S)$ and $\mathcal{A}(S')$ will also be "similar." Then $\mathcal{A}$ doesn't "depend too much" on the particular sample set, and so it's likely to not really overfit too much. (Defining "similar" in different ways will yield different notions of stability.)

In (RLM), the regularizer $R(h)$ "stabilizes" the algorithm. For instance, if you run linear regression on "multicollinear" data (where the $n \times d$ data matrix isn't full rank), there are multiple possible solutions: the algorithm isn't stable. But if you add the regularizer $\lambda \|w\|^2$, then there's always a unique solution that doesn't depend too much on any one input point.

We'll use the following notation for changing single sample points:

$$\text{If } S = (z_1, \dots, z_m), \text{ then } S^{(i \leftarrow z')} = (z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_m).$$

Now, notice the following basically-trivial result:

proposition 1. *For any distribution $\mathcal{D}$ and learning algorithm $\mathcal{A}$,*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m,\,\mathcal{A}}\left[L_{\mathcal{D}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))\right] = \mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m,\,z'\sim\mathcal{D}\\ i\sim\mathrm{Unif}([m]),\,\mathcal{A}}}\left[\ell(\mathcal{A}(S^{(i\leftarrow z')}), z_i) - \ell(\mathcal{A}(S), z_i)\right].$$

On the right-hand side, the first loss term is the generalization loss, since $\mathcal{A}(S^{(i\leftarrow z')})$ doesn't train on $z_i$. The second one is the training error, since $\mathcal{A}(S)$ *does* train on $z_i$.

*Proof.* Splitting up the expectation, the second term has

$$\mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m\\ i\sim\mathrm{Unif}([m]),\,\mathcal{A}}}\left[\ell(\mathcal{A}(S), z_i)\right] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m,\,\mathcal{A}}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\mathcal{A}(S), z_i)\right] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m,\,\mathcal{A}}\left[L_S(\mathcal{A}(S))\right].$$

The first is

$$\mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m,\,z'\sim\mathcal{D}\\ i\sim\mathrm{Unif}([m]),\,\mathcal{A}}}\left[\ell(\mathcal{A}(S^{(i\leftarrow z')}), z_i)\right] = \mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m,\,z'\sim\mathcal{D}\\ i\sim\mathrm{Unif}([m]),\,\mathcal{A}}}\left[\ell(\mathcal{A}(S), z')\right] = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m,\,\mathcal{A}} L_{\mathcal{D}}(\mathcal{A}(S)),$$

where we switched the names of the iid variables $z_i$ and $z'$ in the first equality. □

This motivates what is basically the weakest useful notion of stability:

definition 2. *$\mathcal{A}$ is $\varepsilon(m)$-on-average-replace-one stable* if for all $\mathcal{D}$,

$$\mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^m,\,z'\sim\mathcal{D}\\ i\sim\mathrm{Unif}([m]),\,\mathcal{A}}}\left[\ell(\mathcal{A}(S^{(i\leftarrow z')}), z_i) - \ell(\mathcal{A}(S), z_i)\right] \le \varepsilon(m).$$

We say an algorithm is on-average-replace-one stable if $\lim_{m\to\infty}\varepsilon(m) = 0$ for all $\mathcal{D}$.

Thus, an $\varepsilon(m)$-on-average-replace-one stable algorithm will have small average-case generalization gap $L_{\mathcal{D}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))$: it won't overfit much, on average. Like for uniform convergence, there are bad stable algorithms: consider $\mathcal{A}$ that always returns $x \mapsto 0$ regardless of S. But a stable algorithm that *also* usually fits its training

data, one with small $\mathbb{E}_S L_S(\mathcal{A}(S))$, will have small $\mathbb{E}_S L_{\mathcal{D}}(\mathcal{A}(S))$. This is the notion studied by Chapter 13 of [SSBD]; also see [SSSS10].

Notice that since Proposition 1 is an equality, any algorithm where $\mathbb{E}_S L_{\mathcal{D}}(\mathcal{A}(S))$ is always close to $\mathbb{E}_S L_S(\mathcal{A}(S))$ will necessarily be on-average replace-one stable; this is what I meant by it being the weakest notion. There's also a more commonly considered but much stronger notion of stability:

DEFINITION 3. A learning algorithm $\mathcal{A}$ is $\beta(m)$-*uniformly stable* if for all $m \geq 1$,

$$\sup_{\substack{S \in \mathcal{Z}^m,\, i \in [m] \\ z, z' \in \mathcal{Z}}} \left| \mathop{\mathbb{E}}_{\mathcal{A}} \ell(\mathcal{A}(S^{(i \leftarrow z')}), z) - \mathop{\mathbb{E}}_{\mathcal{A}} \ell(\mathcal{A}(S), z) \right| \leq \beta(m).$$

We say an algorithm is uniformly stable if $\lim_{m \to \infty} \beta(m) = 0$.

*[BE02] defined β as the change from* removing *the ith element, which implies this one with a factor of 2 difference; [EEP05] extended to random algorithms. This version (without the 2) is called* strongly-uniform-replace-one stable *by [SSSS10].*

That is, changing one point in *any* training set gives you a hypothesis that looks almost the same for *any* test point.

If you know differential privacy, this might seem familiar: it turns out that an $(\varepsilon, \delta)$-differentially private algorithm with loss $\ell \in [0, 1]$ is $(e^\varepsilon - 1 + \delta)$-uniformly stable [WLF16, Lemma 23].

If $\mathcal{A}$ is $\beta(m)$-uniformly stable, it's $\varepsilon(m)$-on-average-replace-one stable, just by plugging in the definitions. But the converse is not true. (There are also in-between notions; see [BE02; SSSS10].)

If we do have uniform stability, though, we can get stronger bounds:

THEOREM 4 ([BE02], basically). *Suppose that $\ell(h, z) \in [a, b]$ almost surely. Let $\mathcal{A}$ be $\beta(m)$-uniformly stable. Then, with probability at least $1 - \delta$ over the choice of training points $S \sim \mathcal{D}^m$,*

$$\mathop{\mathbb{E}}_{\mathcal{A}}[L_{\mathcal{D}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))] \leq \beta(m) + (2m\beta(m) + b - a)\sqrt{\frac{1}{2m} \log \frac{1}{\delta}}.$$

The best case for this bound is when $\beta(m) = \mathcal{O}(1/m)$, in which case you get a usual $\mathcal{O}(1/\sqrt{m})$ rate. That turns out to be often the case.

*Proof.* We'll apply McDiarmid's inequality (Theorem 7 of the Rademacher notes) to the function $f(S) = \mathbb{E}_{\mathcal{A}}[L_{\mathcal{D}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))]$.

Since $\beta(m)$-uniformly stable algorithms are $\varepsilon(m)$-on-average-replace-one stable, Proposition 1 implies that $\mathbb{E}_S f(S) \leq \beta(m)$.

For brevity, write $\hat{h}_S$ for $\mathcal{A}(S)$, and $S^i$ for $S^{(i \leftarrow z')}$. Then we have directly that

$$\left| \mathop{\mathbb{E}}_{\mathcal{A}} L_{\mathcal{D}}(\hat{h}_{S^i}) - \mathop{\mathbb{E}}_{\mathcal{A}} L_{\mathcal{D}}(\hat{h}_S) \right| \leq \mathop{\mathbb{E}}_{z \sim \mathcal{D}} \left| \mathop{\mathbb{E}}_{\mathcal{A}} \ell(\hat{h}_{S^i}, z) - \mathop{\mathbb{E}}_{\mathcal{A}} \ell(\hat{h}_S, z) \right| \leq \beta(m).$$

$L_S$ is slightly more complicated, since one evaluation point changes too:

$$\left|\mathop{\mathbb{E}}_{\mathcal{A}} L_{S^i}(\hat{h}_{S^i}) - \mathop{\mathbb{E}}_{\mathcal{A}} L_S(\hat{h}_S)\right| \le \frac{1}{m}\sum_{j\neq i}\left|\mathop{\mathbb{E}}_{\mathcal{A}} \ell(\hat{h}_{S^i}, z_j) - \mathop{\mathbb{E}}_{\mathcal{A}} \ell(\hat{h}_S, z_j)\right| + \frac{1}{m}\left|\mathop{\mathbb{E}}_{\mathcal{A}} \ell(\hat{h}_{S^i}, z') - \mathop{\mathbb{E}}_{\mathcal{A}} \ell(\hat{h}_S, z_i)\right|$$

$$\le \frac{1}{m}\sum_{j\neq i}\beta(m) + \frac{1}{m}(b-a) = \frac{m-1}{m}\beta(m) + \frac{b-a}{m}$$

$$\le \beta(m) + \frac{b-a}{m}.$$

Thus $f$ satisfies the bounded differences condition of McDiarmid with $c_i = 2\beta(m) + \frac{b-a}{m}$. Combining with $\mathbb{E}_S f(S) \le \beta(m)$ gives the desired result. □

*I don't know whether their results hold for random $\mathcal{A}$ or not.*

In fact, even stronger rates have been shown recently, with more complex techniques; Bousquet, Klochkov, and Zhivotovskiy [BKZ20] showed for deterministic $\mathcal{A}$ that, with probability at least $1-\delta$ over the choice of S,

$$L_{\mathcal{D}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S)) \le \text{const}\left(\beta(m)\log(m)\log\frac{1}{\delta} + \frac{b-a}{\sqrt{m}}\sqrt{\log\frac{1}{\delta}}\right) \tag{1}$$

by simplifying and improving techniques of Feldman and Vondrak [FV18; FV19] inspired by techniques from differential privacy. This shows a $\tilde{\mathcal{O}}_p(1/\sqrt{m})$ rate as long as $\beta(m) = \tilde{\mathcal{O}}(1/\sqrt{m})$, much weaker than the $\tilde{\mathcal{O}}(1/m)$ required by Theorem 4.

## 2 Convex Functions

We'll see that (RLM) is often uniformly stable, but to characterize that we'll need various results about *convex functions*. More details are available lots of places; in addition to chapters 12-13 of [SSBD] or Appendix B.2 of [MRT], the classic super-detailed reference is the book of Rockafellar [Roc70], and Boyd and Vandenbreghe [BV04] is also good (and what I learned from).

Most sources assume functions on $\mathbb{R}^d$; we'll assume a separable Hilbert space $\mathcal{X}$, though the statements e.g. that don't use an inner product will also hold for Banach spaces, and so on. For the results about derivatives, you can use a Fréchet derivative, and have a gradient/Hessian analogue. Don't really worry about any of that, you can just think of everything as on $\mathbb{R}^d$.

Definition 5. A *set* $C \subseteq \mathcal{X}$ is convex if for all $x_0, x_1 \in C$ and $\alpha \in [0,1]$, it holds that $(1-\alpha)x_0 + \alpha x_1 \in C$.

Below, we'll use the set $\mathbb{R}\cup\{\infty\}$ a lot. Many of these results hold for the full *extended real line* $\mathbb{R}\cup\{-\infty,\infty\}$, but you often have to exclude $-\infty$ for things to make sense.

It's typical in optimization to, rather than dealing with functions on some restricted domain that's a proper subset of $\mathcal{X}$, instead define $f(x) = \infty$ for $x$ that shouldn't be in the domain. Then $\operatorname{dom} f = \{x \in \mathcal{X} : f(x) < \infty\}$.

Definition 6. A *function* $f : \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ is called

- *convex* if it lies below its chords: for all $x_0, x_1 \in \mathcal{X}$ and $\alpha \in (0,1)$,

$$f((1-\alpha)x_0 + \alpha x_1) \le (1-\alpha)f(x_0) + \alpha f(x_1);$$

- *strictly convex* if this inequality is strict;
- and *m-strongly convex*, for some $m > 0$, if

$$f((1-\alpha)x_0 + \alpha x_1) \leq (1-\alpha)f(x_0) + \alpha f(x_1) - \tfrac{1}{2}m\alpha(1-\alpha)\|x_1 - x_0\|^2.$$

A function is convex if and only if its *epigraph*, $\{(x, r) \in \mathcal{X} \times (\mathbb{R} \cup \{\infty\}) : r \geq f(x)\}$, is a convex set.

An $m$-strongly convex function is $m'$-strongly convex for any $m' < m$; convexity is equivalent to 0-strong convexity, which we don't call strongly convex. $m$-strong convexity implies strict convexity, but the reverse is not true. Likewise, strict convexity implies convexity.

A concave/strictly concave/$m$-strongly concave function is one where $-f$ is convex/strictly convex/$m$-strongly convex.

Any local minimum of a convex function must be a global minimum, since we can connect any two local minima by chords. The set of global minima must be convex, for the same reason. If $f$ is strictly convex, it has only one global minimum.

## 2.1 *First-order conditions*

PROPOSITION 7. *If $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is differentiable on its convex domain,*

- *it is convex iff it lies above its tangents: for all $x, x' \in \operatorname{dom} f$,*

$$f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle.$$

- *it is m-strongly convex iff for all $x, x' \in \mathcal{X}$,*

$$f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle + \tfrac{1}{2}m\left\|x' - x\right\|^2.$$

*Proof.* We'll do this for $m \geq 0$, in which case the $m = 0$ results are for plain convexity.

If $f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2}m\|x' - x\|^2$ for all $x, x'$, then

$$f((1-\alpha)x + \alpha x') \leq (1-\alpha)f(x) + \alpha f(x') - \tfrac{1}{2}m\alpha(1-\alpha)\left\|x' - x\right\|^2$$

$$\frac{1}{\alpha}\Big[f((1-\alpha)x + \alpha x') - f(x)\Big] \leq f(x') - f(x) - \tfrac{1}{2}m(1-\alpha)\left\|x' - x\right\|^2$$

$$\lim_{\alpha \to 0} \frac{f(x + \alpha(x' - x)) - f(x)}{\alpha} \leq f(x') - f(x) - \tfrac{1}{2}m\left\|x' - x\right\|^2,$$

and that limit is exactly the directional derivative given by $\langle \nabla f(x), x' - x \rangle$.

In the other direction, let $x_\alpha = (1-\alpha)x_0 + \alpha x_1$, and note $x_\alpha - x_0 = \alpha(x_1 - x_0)$, $x_\alpha - x_1 = -(1-\alpha)(x_1 - x_0)$. Then

$$\begin{aligned} f(x_\alpha) &\leq f(x_0) + \langle \nabla f(x_\alpha), x_\alpha - x_0 \rangle - \tfrac{1}{2}m\|x_\alpha - x_0\|^2 \\ &= f(x_0) + \alpha\langle \nabla f(x_\alpha), x_1 - x_0 \rangle - \tfrac{1}{2}m\alpha^2\|x_1 - x_0\|^2 \end{aligned}$$

and

$$\begin{aligned} f(x_\alpha) &\leq f(x_1) + \langle \nabla f(x_\alpha), x_\alpha - x_1 \rangle - \tfrac{1}{2}m\|x_\alpha - x_1\|^2 \\ &= f(x_1) - (1-\alpha)\langle \nabla f(x_\alpha), x_1 - x_0 \rangle - \tfrac{1}{2}m(1-\alpha)^2\|x_1 - x_0\|^2. \end{aligned}$$

Adding $1-\alpha$ times the first inequality plus $\alpha$ times the second yields

$$f(x_\alpha) \leq (1-\alpha)f(x_0) + \alpha f(x_1) - \tfrac{1}{2}m\alpha(1-\alpha)(\alpha+1-\alpha)\|x_1 - x_0\|^2 . \qquad \square$$

proposition 8. *If $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is continuously differentiable on its convex domain,*

- *it is convex iff $\forall x, x' \in \operatorname{dom} f$, $\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \geq 0$;*
- *it is $m$-strongly convex iff $\forall x, x' \in \operatorname{dom} f$, $\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \geq m\|x - x'\|^2$.*

This result is important for convex optimization: if $x^*$ is a minimizer, $\nabla f(x^*) = 0$, and so then $m\|x - x^*\|^2 \leq \langle \nabla f(x), x - x^* \rangle \leq \|\nabla f(x)\| \|x - x^*\|$, i.e. $\|x - x^*\| \leq \frac{1}{m}\|\nabla f(x)\|$, and if we know $m > 0$ then the right-hand side is something we can actually measure for any point $x$ and upper-bound how far we can be from the minimizer.

*Proof.* We'll again use $m = 0$ for plain convexity.

If $f$ is convex/$m$-strongly convex, then

$$f(x) \geq f(x') + \langle \nabla f(x'), x - x' \rangle + \tfrac{1}{2}m\left\|x - x'\right\|^2$$
$$f(x') \geq f(x) - \langle \nabla f(x), x - x' \rangle + \tfrac{1}{2}m\left\|x - x'\right\|^2$$

and so

$$f(x) + f(x') \geq f(x') + f(x) + \langle \nabla f(x') - \nabla f(x), x - x' \rangle + m\left\|x - x'\right\|^2 .$$

In the other direction, again using $x_\alpha = (1-\alpha)x_0 + \alpha x_1$ we know that

$$\begin{aligned}
f(x_1) &= f(x_0) + \int_0^1 \langle f(x_\alpha), x_1 - x_0 \rangle \mathrm{d}\alpha \\
&= f(x_0) + \langle \nabla f(x_0), x_1 - x_0 \rangle + \int_0^1 \langle f(x_\alpha) - \nabla f(x_0), x_1 - x_0 \rangle \mathrm{d}\alpha \\
&= f(x_0) + \langle \nabla f(x_0), x_1 - x_0 \rangle + \int_0^1 \frac{1}{\alpha}\langle f(x_\alpha) - \nabla f(x_0), x_\alpha - x_0 \rangle \mathrm{d}\alpha \\
&\geq f(x_0) + \langle \nabla f(x_0), x_1 - x_0 \rangle + \int_0^1 \frac{1}{\alpha} m\|x_\alpha - x_0\|^2 \,\mathrm{d}\alpha \\
&= f(x_0) + \langle \nabla f(x_0), x_1 - x_0 \rangle + m\|x_1 - x_0\|^2 \int_0^1 \alpha \mathrm{d}\alpha \\
&= f(x_0) + \langle \nabla f(x_0), x_1 - x_0 \rangle + \tfrac{1}{2}m\|x_1 - x_0\|^2 .
\end{aligned} \qquad \square$$

## 2.2 *Second-order conditions*

The notation $A \succeq 0$ means that the square matrix (or Hilbert-space operator) A is positive semi-definite; $A \succeq B$ means that $A - B \succeq 0$. Thus $A \succeq mI$ means that all

eigenvalues of A are at least $m$. The notation $\nabla^2 f$ denotes the Hessian, the matrix of all second derivatives. (This is a $\mathcal{F} \to \mathcal{F}$ operator in Hilbert spaces.)

If $f$ is a function on scalars, $\nabla^2 f(x) \geq m\mathrm{I}$ exactly means than $f''(x) \geq m$.

PROPOSITION 9. *If $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is continuously twice-differentiable on its convex domain,*

- *it is convex iff $\forall x \in \operatorname{dom} f$, $\nabla^2 f \geq 0$;*
- *it is $m$-strongly convex iff $\forall x \in \operatorname{dom} f$, $\nabla^2 f \geq m\mathrm{I}$.*

*Proof.* Again use $m = 0$ for the plain convexity case, and $x_\alpha = (1-\alpha)x_0 + \alpha x_1$.

If $f$ is convex / $m$-strongly convex, then using Proposition 8 gives

$$\begin{aligned} m\|x_1 - x_0\|^2 &\leq \langle \nabla f(x_1) - \nabla f(x_0), x_1 - x_0 \rangle \\ &= \left\langle \int_0^1 \nabla^2 f(x_\alpha)(x_1 - x_0)\,\mathrm{d}\alpha, x_1 - x_0 \right\rangle \\ &= \left\langle x_1 - x_0, \left[\int_0^1 \nabla^2 f(x_\alpha)\,\mathrm{d}\alpha\right](x_1 - x_0) \right\rangle \\ 0 &\leq \left\langle x_1 - x_0, \left[\int_0^1 \nabla^2 f(x_\alpha)\,\mathrm{d}\alpha - m\mathrm{I}\right](x_1 - x_0) \right\rangle. \end{aligned}$$

Now, let $x_0$ be any point in the interior of the domain and let $x_1 = x_0 + \varepsilon v$, getting

$$\left\langle \varepsilon v, \left[\int_0^1 \nabla^2 f(x_0 + \varepsilon\alpha v)\,\mathrm{d}\alpha - m\mathrm{I}\right]\varepsilon v \right\rangle \geq 0$$

$$\left\langle v, \left[\int_0^1 \nabla^2 f(x_0 + \varepsilon\alpha v)\,\mathrm{d}\alpha - m\mathrm{I}\right] v \right\rangle \geq 0.$$

As $\varepsilon \to 0$, we have that $\int_0^1 \nabla^2 f(x_0 + \varepsilon\alpha v)\mathrm{d}\alpha \to \nabla^2 f(x_0)$ since $\nabla^2 f$ is continuous. Thus $\langle v, (\nabla^2 f(x) - m\mathrm{I})v\rangle \geq 0$ for all $x$ in the interior of the domain and all $v$. This is exactly the condition that $\nabla^2 f(\mathrm{X}) \geq m\mathrm{I}$.

For the other direction, we have that

$$\begin{aligned} f(x') &= f(x) + \langle \nabla f(x), x' - x\rangle + \int_0^1 \int_0^\alpha \langle x' - x, \nabla^2 f(x_\tau)(x' - x)\rangle \mathrm{d}\tau\,\mathrm{d}\alpha \\ &\geq f(x) + \langle \nabla f(x), x' - x\rangle + \int_0^1 \int_0^\alpha m\left\|x' - x\right\|^2 \mathrm{d}\tau\,\mathrm{d}\alpha \\ &= f(x) + \langle \nabla f(x), x' - x\rangle + \tfrac{1}{2} m\left\|x' - x\right\|^2 \end{aligned}$$

since $\int_0^\alpha d\tau = \alpha$, and $\int_0^1 \alpha \, d\alpha = \frac{1}{2}$. □

### 2.3 *Properties*

proposition 10. *If $f$, $g$, and $f_y$ for all $y \in \mathcal{Y}$ are all convex functions, then so are*

- *$\alpha f$ for any $\alpha \geq 0$;*
- *$f + g$, or more generally $\int f_y dw(y)$ if $w$ is any (nonnegative) measure on $\mathcal{Y}$;*
- *$x \mapsto f(Ax + b)$ for any $A, b$;*
- *$x \mapsto g(f(x))$ if $g : \mathbb{R} \to \mathbb{R}$ is also nondecreasing;*
- *$x \mapsto \max(f(x), g(x))$, or more generally $x \mapsto \sup_{y \in \mathcal{Y}} f_y(x)$;*
- *$x \mapsto \inf_{y \in \mathcal{Y}} f(x, y)$ if $f(x, y)$ is convex in $(x, y)$, and $\mathcal{Y}$ is a nonempty convex set.*

The proofs are mostly straightforward, and omitted here.

Similarly, the sum of an $m$-strongly convex and an $m'$-strongly convex function is $(m + m')$-strongly convex, and the sum of an $m$-strongly convex function with a convex function is $m$-strongly convex. Scaling an $m$-strongly convex function by $\alpha > 0$ gives you an $m\alpha$-strongly convex function.

Notice that the square loss, hinge loss, and logistic loss are all convex functions of the function $h$.

theorem 11 (Jensen's inequality). *If $f : \mathcal{X} \to \mathbb{R}$ is convex and $X$ a random variable on $\mathcal{X}$ such that the expectations exist, $f(\mathbb{E} X) \leq \mathbb{E} f(X)$.*

We also have that $\|h\|^2$ is 2-strongly convex: for instance, its gradient is $2h$ and so its Hessian is $2I$, or you can use directly that

$$\begin{aligned}
\|(1-\alpha)h + \alpha g\|^2 + \tfrac{1}{2} \cdot 2 \cdot \alpha(1-\alpha)\|h - g\|^2 & \\
&= (1-\alpha)^2 \|h\|^2 + \alpha^2 \|g\|^2 + 2\alpha(1-\alpha)\langle h, g\rangle \\
&\quad + \alpha(1-\alpha)\|h\|^2 + \alpha(1-\alpha)\|g\|^2 - 2\alpha(1-\alpha)\langle h, g\rangle \\
&= (1 - \alpha + \alpha)(1-\alpha)\|h\|^2 + \alpha(\alpha + 1 - \alpha)\|g\|^2 \\
&= (1-\alpha)\|h\|^2 + \alpha\|g\|^2 .
\end{aligned}$$

Thus $\frac{1}{2}\|h\|^2$ is 1-strongly convex.

## 3 CONVEX RLM

*If it's strongly convex with some different $m$, just scale it and inversely scale $\lambda$.*

Recall regularized loss minimization (RLM), and suppose $h \mapsto \ell(h, z)$ is convex for each $z$, and $R(h)$ is 1-strongly convex. Then $f_S(h) = L_S(h) + \lambda R(h)$, the sum of a convex function and a $\lambda$-strongly convex function, is $\lambda$-strongly convex. Let $\mathcal{A}(S)$ denote $\arg\min_{h \in \mathcal{H}} f_S(h)$; since $f_S$ is strongly convex, it has a unique minimizer.

Now, notice that for any $h \in \mathcal{H}$, we have that

$$f_S(h) = f_{S^{(i \leftarrow z')}}(h) - \frac{1}{m}\ell(h, z') + \frac{1}{m}\ell(h, z_i).$$

Thus for any $h, g \in \mathcal{H}$,

$$f_S(h) - f_S(g) = f_{S^{(i\leftarrow z')}}(h) - f_{S^{(i\leftarrow z')}}(g) + \frac{\ell(h, z_i) - \ell(g, z_i)}{m} + \frac{\ell(g, z') - \ell(h, z')}{m}.$$

Plugging in $h = \hat{h}^i = \mathcal{A}(S^{(i\leftarrow z')})$ and $g = \hat{h} = \mathcal{A}(S)$, since $\hat{h}^i$ minimizes $f_{S^{(i\leftarrow z')}}$, we get that

$$f_S(\hat{h}^i) - f_S(\hat{h}) \le \frac{\ell(\hat{h}^i, z_i) - \ell(\hat{h}, z_i)}{m} + \frac{\ell(\hat{h}, z') - \ell(\hat{h}^i, z')}{m}.$$

Noting also that $\nabla f_S(\hat{h}) = 0$, Proposition 7 implies that $f_S(\hat{h}^i) - f_S(\hat{h}) \ge \frac{1}{2}\lambda \left\|\hat{h}^i - \hat{h}\right\|^2$. Thus we've shown that

$$\frac{\lambda}{2}\left\|\hat{h}^i - \hat{h}\right\|^2 \le \frac{\ell(\hat{h}^i, z_i) - \ell(\hat{h}, z_i)}{m} + \frac{\ell(\hat{h}, z') - \ell(\hat{h}^i, z')}{m}.$$

### 3.1 *Lipschitz loss*

If we further assume that $h \mapsto \ell(h, z)$ is $\rho$-Lipschitz for each $z$, i.e. $|\ell(h, z) - \ell(h', z)| \le \rho\|h - h'\|$ for all $z$, then we get that

$$\frac{\lambda}{2}\left\|\hat{h}^i - \hat{h}\right\|^2 \le \frac{2\rho}{m}\left\|\hat{h}^i - \hat{h}\right\|,$$

and hence

$$\left\|\hat{h}^i - \hat{h}\right\| \le \frac{4\rho}{\lambda m}.$$

Using the Lipschitz property again,

$$\left|\ell(\hat{h}^i, z) - \ell(\hat{h}, z)\right| \le \rho\left\|\hat{h}^i - \hat{h}\right\| \le \frac{4\rho^2}{\lambda m}$$

for any $z$, and we've shown $\frac{4\rho^2}{\lambda m}$-uniform stability. Plugging into Theorem 4, we get

PROPOSITION 12. *Suppose that $h \mapsto \ell(h, z)$ is $\rho$-Lipschitz and convex for each $z \in \mathcal{Z}$. Let $h \mapsto \lambda R(h)$ be $\lambda$-strongly convex. Then (RLM) satisfies that*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \le \frac{4\rho^2}{\lambda m}.$$

*If we further have that $\ell(h, z) \in [a, b]$ for all $h$ and $\mathcal{D}$-almost all $z$, (RLM) satisfies that, with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$,*

$$L_{\mathcal{D}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S)) \le \frac{4\rho^2}{\lambda m} + \left(\frac{8\rho^2}{\lambda} + b - a\right)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}.$$

### 3.2 *Smooth, nonnegative loss*

In optimization, a $\beta$-*smooth function* is one whose gradient is $\beta$-Lipschitz. It's possible, though messier, to show a similar result for nonnegative $\beta$-smooth functions: in the end, if you also assume that $\lambda \ge 2\beta/m$, you get (see [SSBD, Section 13.3.2]) that

$$\mathop{\mathbb{E}}_{S,z'}\left[L_S(\mathcal{A}(S^{(i\leftarrow z')})) - L_S(\mathcal{A}(S))\right] \le \frac{48\beta}{\lambda m}\mathop{\mathbb{E}}_{S}[L_S(\mathcal{A}(S))].$$

If $R$ is nonnegative and there is some $h_0$ with $R(h_0) = 0$, e.g. the zero predictor when

$R(h) = \|h\|^2$, and we also have that $\ell(h_0, z) \le b$, say because the loss is always less than $b$, then we know that

$$L_S(\mathcal{A}(S)) \le L_S(\mathcal{A}(S)) + \lambda R(\mathcal{A}(S)) \le L_S(h_0) + \lambda R(h_0) \le b,$$

in which case we can upper-bound $\mathbb{E}_S L_S(\mathcal{A}(S))$ by $b$.

## 4 Fitting-stability trade-off

Proposition 12 shows that $L_{\mathcal{D}}(\mathcal{A}(S))$ is closer to $L_S(\mathcal{A}(S))$ as $\lambda$ grows. But $L_S(\mathcal{A}(S))$ will also get worse for bigger $\lambda$, since $\mathcal{A}$ cares more and more about having "simpler" $h$ than minimizing $L_S$. How can we find the best $\lambda$ to trade off between them?

For any fixed $h^*$, if R is nonnegative and 1-strongly convex, we have that

$$L_S(\mathcal{A}(S)) \le L_S(\mathcal{A}(S)) + \lambda R(h) \le L_S(h^*) + \lambda R(h^*),$$

and hence

$$\mathop{\mathbb{E}}_S L_S(\mathcal{A}(S)) \le L_{\mathcal{D}}(h^*) + \lambda R(h^*). \tag{2}$$

### 4.1 *Expected loss*

So, using (2) and Propositions 1 and 12,

$$\begin{aligned}\mathop{\mathbb{E}}_S L_{\mathcal{D}}(\mathcal{A}(S)) &\le L_{\mathcal{D}}(h^*) + \lambda R(h^*) + \mathop{\mathbb{E}}_S[L_{\mathcal{D}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))] \\ &\le L_{\mathcal{D}}(h^*) + \lambda R(h^*) + \frac{4\rho^2}{\lambda m} \end{aligned} \tag{3}$$

if the loss is convex and $\rho$-Lipschitz, and R is nonnegative and 1-strongly convex.

We now have two routes we can take from this bound.

#### 4.1.1 Choosing $\lambda$

Because our canonical regularizer is $R(h) = \frac{1}{2}\|h\|^2$, let's compare our result to any $h^*$ satisfying $R(h^*) \le \frac{1}{2}B^2$. We then have

$$\mathop{\mathbb{E}}_S L_{\mathcal{D}}(\mathcal{A}(S)) \le L_{\mathcal{D}}(h^*) + \frac{1}{2}\lambda B^2 + \frac{4\rho^2}{\lambda m}. \tag{4}$$

Using that $\alpha x + \beta/x$ is minimized at $x = \sqrt{\beta/\alpha}$ with value $2\sqrt{\alpha\beta}$, we can minimize this bound by picking $\lambda = \frac{\rho}{B}\sqrt{\frac{8}{m}}$ to get

$$\mathop{\mathbb{E}}_S L_{\mathcal{D}}(\mathcal{A}(S)) \le \inf_{h^*: R(h^*) \le \frac{1}{2}B^2} L_{\mathcal{D}}(h^*) + \rho B\sqrt{\frac{8}{m}}.$$

*The class of problems is also sometimes described in terms of Bregman divergences; see e.g. Chapter 14 of [MRT].*

**Definition 13.** A learning problem is *convex*, $\rho$-*Lipschitz*, B-*bounded* if the hypothesis set $\mathcal{H}$ is convex, bounded as $R(h) \le \frac{1}{2}B^2$ for all $h \in \mathcal{H}$ for some strongly convex function R, and the loss functions $h \mapsto \ell(h, z)$ are convex and $\rho$-Lipschitz for each $z \in \mathcal{Z}$.

Regularized loss minimization can therefore learn (in the sense of expected loss) any convex-Lipschitz-bounded learning problem. (Section 12.2.1 of [SSBD] gives

examples of problems which are not learnable if they're convex but not Lipschitz and/or not bounded.)

You can do a similar thing for the case where the loss is nonnegative and smooth instead of Lipschitz [SSBD, Corollaries 13.10 and 13.11].

### 4.1.2 Fixing $\lambda$

Instead of fixing some B and picking $\lambda$ accordingly, we can instead ask: what happens if we run the algorithm with any particular choice of $\lambda$? We minimized (4) in terms of $\lambda$, but it's equivalent to take $\lambda$ as given and minimize the bound in terms of B, in which case we find $B = \frac{\rho}{\lambda}\sqrt{\frac{8}{m}}$ and

$$\mathop{\mathbb{E}}_{S} L_{\mathcal{D}}(\mathcal{A}(S)) \le \inf_{h^* : R(h^*) \le \frac{1}{2}\left(\frac{\rho}{\lambda}\sqrt{\frac{8}{m}}\right)^2} L_{\mathcal{D}}(h^*) + \frac{8\rho^2}{\lambda m}.$$

This is exactly the same analysis, but tells us what happens with different choices of $\lambda$: for instance, if we pick a constant $\lambda$ as $m$ grows, then we only compete with simpler and simpler hypotheses as we see more data (not ideal).

We need $\frac{1}{\lambda m} \to 0$ for the second term to go to zero, i.e. we need $\lambda = \omega(1/m)$. If we also want to eventually compete with *any* possible predictor, we need $\frac{1}{\lambda\sqrt{m}} \to \infty$, i.e. $\lambda = o(1/\sqrt{m})$. Thus, consider picking $\lambda \propto m^{-\gamma}$ for any $\gamma \in (1/2, 1)$: the second term becomes $\mathcal{O}(m^{\gamma-1})$.

Note that this means boundedness isn't actually required for learnability: convexity and Lipschitzness of the loss is enough. With boundedness, however, we know that $\inf_{h* : R(h^*) \le \frac{1}{2}\left(\frac{\rho}{\lambda}\sqrt{\frac{8}{m}}\right)^2} L_{\mathcal{D}}(h^*)$ will become exactly $\inf_{h^* \in \mathcal{H}} L_{\mathcal{D}}(h^*)$ for $m$ bigger than some threshold. Without boundedness, it might be that $\inf_{h^* \in \mathcal{H}} L_{\mathcal{D}}(h^*)$ is not actually achieved, only approached as $R(h^*) \to \infty$; this is actually the case for logistic regression of separable data. To get a final bound, then, we would need to know the rate at which $\inf_{h* : R(h^*) \le \frac{1}{2}\left(\frac{\rho}{\lambda}\sqrt{\frac{8}{m}}\right)^2} L_{\mathcal{D}}(h^*)$ approaches its asymptote as $m$ grows.

## 4.2 *High-probability bound*

To get high-probability bounds, also assume that $\ell \in [a, b]$.

We showed in the proof of Theorem 4 that $L_{\mathcal{D}}(\mathcal{A}(S))$ satisfies bounded differences with $c_i = \beta(m)$. We then have that $L_{\mathcal{D}}(\mathcal{A}(S))$ is close to $\mathbb{E}_S L_{\mathcal{D}}(\mathcal{A}(S))$, which we just upper-bounded; we don't actually need to apply Theorem 4 directly, and doing so wouldn't be any tighter. This tells us that

$$L_{\mathcal{D}}(\mathcal{A}(S)) \le \mathop{\mathbb{E}}_{S} L_{\mathcal{D}}(\mathcal{A}(S)) + \frac{4\rho^2}{\lambda m}\sqrt{\frac{m}{2}\log\frac{1}{\delta}},$$

and plugging in (4) gives for any fixed $h^*$, it holds with probability at least $1 - \delta$ that

$$L_{\mathcal{D}}(\mathcal{A}(S)) \le L_{\mathcal{D}}(h^*) + \lambda R(h^*) + \frac{4\rho^2}{\lambda m} + \frac{4\rho^2}{\lambda\sqrt{m}}\sqrt{\frac{1}{2}\log\frac{1}{\delta}}. \tag{5}$$

We can see that the choice $\lambda \propto 1/\sqrt{m}$, which was optimal in the average-case analysis of Section 4.1.1, will now give us a *constant* upper bound. That's no good. The problematic term is $\beta(m)\sqrt{m}$, the same as in Theorem 4, but using $\lambda \propto 1/\sqrt{m}$ means that $\beta(m) = \mathcal{O}(\frac{1}{\lambda m})$ becomes $\mathcal{O}(1/\sqrt{m})$.

We'll thus need a larger $\lambda$. Assuming $R(h^*) \leq \frac{1}{2}B^2$, we can minimize (5) when

$$\lambda = \sqrt{\frac{2}{B^2} \cdot 4\rho^2 \left( \frac{1}{m} + \sqrt{\frac{1}{2m} \log \frac{1}{\delta}} \right)} = \frac{\rho}{B} \sqrt{8 \left( \frac{1}{m} + \sqrt{\frac{1}{2m} \log \frac{1}{\delta}} \right)},$$

giving a bound of

$$L_{\mathcal{D}}(\mathcal{A}(S)) - L_{\mathcal{D}}(h^*) \leq B\rho \sqrt{8 \left( \frac{1}{m} + \sqrt{\frac{1}{2m} \log \frac{1}{\delta}} \right)}.$$

Notice that this upper bound on the suboptimality is bigger than $B\rho \left( \frac{32}{m} \log \frac{1}{\delta} \right)^{1/4}$; that $m^{-1/4}$ rate is much slower than the $m^{-1/2}$ rate for the mean!

Using the sharper bound of (1) instead gives a much better result:

$$L_{\mathcal{D}}(\mathcal{A}(S)) - L_{\mathcal{D}}(h^*) = \mathcal{O}\left( \lambda B^2 + \frac{\rho^2}{\lambda m} + \frac{\rho^2}{\lambda m} \log m \log \frac{1}{\delta} + \frac{b-a}{\sqrt{m}} \sqrt{\log \frac{1}{\delta}} \right).$$

This upper bound is minimized by $\lambda = \Theta\left( \frac{\rho}{B} \sqrt{\frac{1 + \log m \log \frac{1}{\delta}}{m}} \right) = \tilde{\Theta}\left( \frac{\rho}{B\sqrt{m}} \right)$, giving

$$L_{\mathcal{D}}(\mathcal{A}(S)) - L_{\mathcal{D}}(h^*) = \mathcal{O}\left( B\rho \sqrt{\frac{1 + \log m \log \frac{1}{\delta}}{m}} + \frac{b-a}{\sqrt{m}} \sqrt{\log \frac{1}{\delta}} \right) = \tilde{\mathcal{O}}_p\left( \frac{\max(B\rho, b-a)}{\sqrt{m}} \right).$$

This is a far more satisfying final bound, and notice that the choice of $\lambda$ is similar (up to log factors) to the case from Section 4.1.1 where we only analyzed the mean.

Either of these analyses shows that convex-Lipschitz-bounded learning problems are also learnable by RLM with high probability.

A similar, but messier, analysis should also work for the case of a nonnegative $\beta$-smooth loss.

## 5 MORE

Other algorithms than RLM are also stable. The theory originated out of analyses of "local learning rules" like $k$-nearest neighbour, which are on-average-replace-one stable (and "hypothesis stable") but not uniformly stable. You can also show that (stochastic) gradient descent is uniformly stable for convex problems, and even get some results in non-convex settings [HRS15; FV19]. And, as mentioned, any differentially private algorithm is automatically uniformly stable (with the randomized variant).

Stability is also very useful for analyzing cross-validation, especially leave-one-out cross-validation, which is discussed by many of the papers here [BE02; SSSS10].

Shalev-Shwartz et al. [SSSS10] also have thorough accounting of when stability is necessary, or not, to be able to learn different kinds of problems, and Wang, Lei, and Fienberg [WLF16] a similar accounting of the relationship to privacy.

## REFERENCES

[BE02] Olivier Bousquet and André Elisseeff. "Stability and Generalization." *Journal of Machine Learning Research* 2 (2002), pages 499–526.

[BKZ20] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. "Sharper Bounds for Uniformly Stable Algorithms." *Conference on Learning Theory*. 2020.

[BV04] Stephen Boyd and Lieven Vandenbreghe. *Convex Optimization*. Cambridge University Press, 2004.

[EEP05] Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. "Stability of Randomized Learning Algorithms." *Journal of Machine Learning Research* 6.3 (2005), pages 55–79.

[FV18] Vitaly Feldman and Jan Vondrak. "Generalization Bounds for Uniformly Stable Algorithms." *Advances in Neural Information Processing Systems*. Volume 31. 2018. arXiv: 1812.09859.

[FV19] Vitaly Feldman and Jan Vondrak. "High probability generalization bounds for uniformly stable algorithms with nearly optimal rate." *Conference on Learning Theory*. 2019.

[HRS15] Moritz Hardt, Benjamin Recht, and Yoram Singer. "Train faster, generalize better: Stability of stochastic gradient descent." *Advances in Neural Information Processing Systems*. 2015. arXiv: 1509.01240.

[MRT] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talkwalkar. *Foundations of Machine Learning*. 2nd edition. MIT Press, 2018.

[Roc70] R. Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[SSBD] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[SSSS10] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. "Learnability, Stability and Uniform Convergence." *Journal of Machine Learning Research* 11 (2010), pages 2635–2670.

[WLF16] Yu-Xiang Wang, Jing Lei, and Stephen E. Fienberg. "Learning with Differential Privacy: Stability, Learnability and the Sufficiency and Necessity of ERM Principle." *Journal of Machine Learning Research* 17.183 (2016), pages 1–40.