Proceedings of Machine Learning Research vol 65:1–15, 2017

A Second-order Look at Stability and Generalization

Andreas Maurer AM@ANDREAS-MAURER.EU
Adalbertstr 55, D-80799 Munich, Germany

Abstract

Using differentiability assumptions on the loss function and a concentration inequality for bounded second order differences it is shown that the generalization error for classification with L2 regularisation obeys a Bernstein-type inequality.

Keywords: generalisation, stability, concentration, Bernstein inequality

1. Introduction and Main Results

This work studies some properties of the function

$$g(\mathbf{x}) = \arg\min_{w \in H} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\langle x_i, w\rangle\right) + \lambda \|w\|^2, \tag{1}$$

where $(H, \langle .,. \rangle, \|.\|)$ is a real Hilbert space, $\mathbf{x}$ is a sample of vectors $x_1, ..., x_n$ drawn independently from a distribution μ on the unit ball $\mathbb{B}$ of H, the non-negative real loss function ℓ is assumed to be convex and to satisfy $\ell(0) = 1$, and the regularization parameter λ satisfies $0 < \lambda < 1$. For clarity we consider classification only, with labels being absorbed in the data vectors. The first term on the right hand side of (1) is called the empirical risk

$$\hat{L}(\mathbf{x}) = \frac{1}{n} \sum_{i} \ell\left(\langle x_i, g(\mathbf{x})\rangle\right).$$

A central question in learning theory is, to which extent the empirical risk can be used to bound the true risk $L(\mathbf{x}) = E_{x\sim\mu}\left[\ell\left(\langle x, g(\mathbf{x})\rangle\right)\right]$. One wishes to bound the generalization error

$$\Delta(\mathbf{x}) = L(\mathbf{x}) - \hat{L}(\mathbf{x})$$

in terms of the sample size n, the regularization parameter λ and properties of the loss function ℓ. Because of the dependence on the random sample $\mathbf{x}$ such bounds are necessarily probabilistic with given violation probability δ.

One method to tackle this problem is to recognise the implicit constraint $\|w\| \leq \lambda^{-1/2}$ and to reformulate (1) as a problem of empirical risk minimization over some hypothesis space whose capacity is then controlled with empirical process theory (Anthony and Bartlett (1999), Bartlett and Mendelson (2002) and others). Without additional information the resulting bounds on the generalization error are worst-case bounds over the hypothesis space and of order $1/\sqrt{n}$.

Another line of thought refrains from analysis of a hypothesis space and focuses on the stability properties of the function g. If the loss function has appropriate Lipschitz properties then bounds on the maximal dependence of $g(\mathbf{x})$ on variations in any data point x_i lead to high probability bounds on $\Delta(\mathbf{x})$. In a seminal paper Bousquet and Elisseeff (2002) prove the following result.

Theorem 1 *If ℓ is Lipschitz, there are finite quantities γ_1 and γ_2, depending on ℓ and λ only, such that for every $\delta > 0$ with probability at least $1-\delta$ in $\mathbf{x} \sim \mu^n$*

$$\Delta(\mathbf{x}) \leq \frac{\gamma_1(\ell,\lambda)}{n} + \sqrt{\frac{\gamma_2(\ell,\lambda)\ln(1/\delta)}{n}}.$$

Here the first term on the right hand side is a bound on the expectation $E[\Delta(\mathbf{x})]$, while the second term is a bound on the estimation difference $\Delta(\mathbf{x}) - E[\Delta(\mathbf{x})]$, which is obtained from the bounded difference inequality (McDiarmid (1998), Boucheron et al. (2013)).

The derivation of Theorem 1 in Bousquet and Elisseeff (2002) is elegant and general, but the bound leaves no room for potentially beneficial properties of the distribution μ. In some sense it still is a worst-case bound. Here we propose an alternative bound in the case that ℓ has a third derivative.

Theorem 2 *If $\ell \in C^3(\mathbb{R})$, then there are finite quantities α_1 and α_2, depending on ℓ and λ only, such that for every $\delta \in (0, 1/e)$ with probability at least $1-\delta$ in $\mathbf{x} \sim \mu^n$*

$$\Delta(\mathbf{x}) \leq \frac{\alpha_1(\ell,\lambda)}{n} + \sqrt{2\sigma^2(\Delta)\ln(1/\delta)} + \frac{\alpha_2(\ell,\lambda)\ln(1/\delta)}{n},$$

where $\sigma^2(\Delta)$ is the variance of Δ

$$\sigma^2(\Delta) = E_{\mathbf{x}\sim\mu^n}\left[\left(\Delta(\mathbf{x}) - E_{\mathbf{x}'\sim\mu^n}\left(\Delta(\mathbf{x}')\right)\right)^2\right],$$

and the functionals $\alpha_i : C^3(\mathbb{R}) \times (0,1) \to \mathbb{R}^+$ are given by

$$\begin{aligned}\alpha_1(\ell,\lambda) &= \frac{2\max\{1,c',c''\}^2}{\lambda^{3/2}} \\ \alpha_2(\ell,\lambda) &= \left(\frac{68}{\lambda^3} + \frac{24c'''}{\lambda^4}\right)\max\{1,c',c''\}^3,\end{aligned}$$

with $c^{(i)} := \sup_{|t|\leq(1/\lambda)^{1/2}} \left|\ell^{(i)}(t)\right|$ and $\ell^{(i)}$ being the i-th derivative of ℓ.

This is our main result, and before describing its merits we list some of its shortcomings, accompanied by respective apologies.

1. In practice λ is often chosen to decrease with increasing sample size as $\lambda \approx n^{-p}$. If we want the last term to go to zero, we are forced to choose $p < 1/4$, or, if $c''' = 0$ as for the square loss, $p < 1/3$. These exponents being impractical, the bound should really only be used to study behavior of the algorithm for $\lambda > 0$ fixed, which we shall assume in the sequel. This might be an artifact of the proof, but poor dependence on λ might also be a general problem of the stability approach, as also noted in the discussion section of Bousquet and Elisseeff (2002), where we are forced to $p < 1/2$ or $p < 1/3$.

2. The differentiability requirement is a severe limitation on the scope of the result. It excludes the hinge loss, for example. As it stands, the only frequently used loss functions to which the result applies are the square and the logistic loss. The assumption $\ell \in C^3$ is necessary to bound second order differences of $\Delta(\mathbf{x})$ by differentiation, but there might be some other method. Some hope is inspired by the success of Bousquet and Elisseeff (2002) in bounding first order variations

without the assumption of differentiability. Whether their method can be extended to second order differences, or not, remains an open question.

3. The large constants in the definitions of the α_i and the way in which the $c^{(i)}$ are combined appear excessive. These expressions are not optimal, not even in the context of the given method of proof, but they were deliberately chosen for a compact and readable appearance.

4. The restriction to classification, with labels being absorbed in the data vectors, avoids additional but elementary complications and was made for reasons of presentation. The reader who studies the proof will verify that the method can be extended to regression.

As a positive side of Theorem 2 first note that, as n increases, always with λ fixed, the $O(1/n)$ terms decay rapidly until the bound is dominated by the variance term, a behavior reminiscent of Bernstein's inequality. In fact, if we neglect the $O(1/n)$ terms and solve for δ, we get, for $t > 0$,

$$\Pr\{\Delta > t\} \lesssim \exp\left(\frac{-t^2}{2\sigma^2(\Delta)}\right),$$

which is the tailbound for a centered normal variable with variance $\sigma^2(\Delta)$. It will become apparent from the proof, that similar concentration properties hold for L, $\hat{L}$ and, in a weak sense, for the vector valued function g.

Next we argue that the term $2\sigma^2(\Delta)$ in Theorem 2 can never be larger than the term $\gamma_2(\ell,\lambda)/n$ in Theorem 1, in fact, it can never be larger than anything derived from the bounded difference inequality. To show this we introduce some notation, which will be useful in the sequel.

With $\mathcal{A}_n$ we denote the algebra of bounded real-valued measurable functions on $\mathbb{B}^n$ (the space of samples). For $y, y' \in \mathbb{B}$ and $k \in \{1, ..., n\}$ we define the difference operator $D^k_{y,y'} : \mathcal{A}_n \to \mathcal{A}_n$ for $f \in \mathcal{A}_n$ by

$$\left(D^k_{y,y'} f\right)(\mathbf{x}) = f(x_1, ..., x_{k-1}, y, x_{k+1}, ..., x_n) - f\left(x_1, ..., x_{k-1}, y', x_{k+1}, ..., x_n\right).$$

Note that $\left(D^k_{y,y'} f\right)(\mathbf{x})$ is independent of x_k.

The bounded difference inequality (McDiarmid (1998), Boucheron et al. (2013)) asserts, that for any $f \in \mathcal{A}_n$ and $t > 0$

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-2t^2}{\sup_{\mathbf{x}\in\mathbb{B}^n} \sum_k \sup_{y,y'\in\mathbb{B}} \left(D^k_{y,y'} f(\mathbf{x})\right)^2}\right).$$

In Bousquet and Elisseeff (2002), for $f = \Delta$, the expression in the denominator of the exponent is bounded by $2\gamma_2(\ell,\lambda)/n$ to prove Theorem 1. Now we define the k-th conditional variance as the nonlinear operator $\sigma^2_k : \mathcal{A}_n \to \mathcal{A}_n$

$$\sigma^2_k(f)(\mathbf{x}) = \frac{1}{2} E_{(y,y')\sim\mu^2}\left(D^k_{y,y'} f(\mathbf{x})\right)^2.$$

This is just the variance of f in the k-th variable, conditional on all the other variables. We also define the nonlinear operator $\Sigma^2 : \mathcal{A}_n \to \mathcal{A}_n$ by $\Sigma^2(f) = \sum_k \sigma^2_k(f)$. It is common knowledge

that the variance of a bounded random variable is no larger than a quarter of the square of its range (see Lemma 8 in Appendix A), so

$$\sup_{\mathbf{x}\in\mathbb{B}^n}\Sigma^2\left(\Delta\right)\left(\mathbf{x}\right)\leq\frac{1}{4}\sup_{\mathbf{x}\in\mathbb{B}^n}\sum_k\sup_{y,y'\in\mathbb{B}}\left(D^k_{y,y'}\Delta\left(\mathbf{x}\right)\right)^2\leq\frac{\gamma_2\left(\ell,\lambda\right)}{2n}.$$

The Efron Stein inequality (Efron and Stein (1981), Steele (1986), Boucheron et al. (2013)) on the other hand, asserts that $\sigma^2\left(f\right)\leq E\left[\Sigma^2\left(f\right)\right]$ for $f\in\mathcal{A}_n$, so

$$2\sigma^2\left(\Delta\right)\leq 2E\left[\Sigma^2\left(\Delta\right)\right]\leq 2\sup_{\mathbf{x}\in\mathbb{B}^n}\Sigma^2\left(\Delta\right)\left(\mathbf{x}\right)\leq\frac{\gamma_2\left(\ell,\lambda\right)}{n}. \tag{2}$$

So, once the $O\left(1/n\right)$-terms are out of the way, the bound in Theorem 2 can never be larger than the one in Theorem 1, but, depending on μ, it may be significantly smaller. In fact, suppose that one can show (by appropriate assumptions on the distribution), that for some constant c, $p\in[1/2,1]$ and all $\delta>0$ with probability at most $1-\delta$ we have $\Delta\leq cn^{-p}\ln\left(1/\delta\right)$. Then, since Δ can be shown to be bounded and letting $\delta=n^{-2p}$, it easily follows that $\sigma^2\left(\Delta\right)\leq Cn^{-2p}\ln\left(n\right)$ for some other constant C. In this sense Theorem 2 inherits any fast-rate bound up to a logarithmic factor. The case of finite dimensional H together with the square loss then furnishes an example.

In the proof of Theorem 2 the crucial property of the function Δ is *first- and second-order stability*: For some C_1 and C_2 (depending only on ℓ and λ) we have

$$\begin{aligned}\max_k\sup_{\mathbf{x}\in\mathbb{B}^n,y,y'\in\mathbb{B}}D^k_{y,y'}\Delta\left(\mathbf{x}\right)&\leq C_1/n\\ \max_{k\neq l}\sup_{\mathbf{x}\in\mathbb{B}^n,z,z',y,y'\in\mathbb{B}}D^l_{z,z'}D^k_{y,y'}\Delta\left(\mathbf{x}\right)&\leq C_2/n^2.\end{aligned}$$

So, in addition to uniform stability as in Bousquet and Elisseeff (2002), we require that, for any training sample $\mathbf{x}$, any variation of Δ, which is induced by modification of one data point, can not change by more than C_2/n^2, if we modify another data point. Similar results to Theorem 2 can be obtained for any algorithm for which both these stability requirements can be verified.

Related to Theorem 2 is the following result, which is in some sense intermediate between Theorem 1 and Theorem 2.

Theorem 3 *If $\ell\in C^2$ then for every $\delta>0$ with probability at least $1-\delta$ in $\mathbf{x}\sim\mu^n$*

$$\begin{aligned}\Delta\left(\mathbf{x}\right)&\leq\frac{\alpha_1\left(\ell,\lambda\right)}{n}+\sqrt{2\sup_{\mathbf{x}\in\mathbb{B}^n}\Sigma^2\left(\Delta\right)\left(\mathbf{x}\right)\ln\left(1/\delta\right)}+\frac{2\alpha_1\left(\ell,\lambda\right)\ln\left(1/\delta\right)}{n}\\&\leq\alpha_1\left(\ell,\lambda\right)\left(\frac{1}{n}+3\sqrt{\frac{2\sigma^2_{x\sim\mu}\left(x\right)\ln\left(1/\delta\right)}{n}}+\frac{2\ln\left(1/\delta\right)}{n}\right),\end{aligned}$$

where

$$\sigma^2_{x\sim\mu}\left(x\right)=\frac{1}{2}E_{(x,x')\sim\mu^2}\left\|x-x'\right\|^2$$

is the variance of the identity map in $\mathbb{B}$ under the law μ.

For small n the first inequality above is much stronger than Theorem 2, since α_1 has a better dependence on λ than α_2. For large n it becomes weaker than Theorem 2 and stronger than Theorem 1 because of (2). In any case: if $\ell \in C^3$ we can always take the smallest of the three bounds. The second inequality of Theorem 3 may be of independent interest.

The function (1) has been studied extensively (see e.g. Poggio and Girosi (1990), Cucker and Smale (2002), Caponnetto and De Vito (2007)) from the perspectives of machine learning and inverse-problem theory, with frequent focus on the square loss, but the author is not aware of distribution dependent results of comparable simplicity to Theorem 2. The present work is motivated by a renewed interest in the stability approach because of its promise for the analysis of complex learning machines (see Hardt et al. (2015)). The most challenging problem here is to improve on the functions α_1 and α_2, both in their dependence on λ and on the differentiability assumptions.

The next section gives proofs of Theorem 2 and Theorem 3. Appendix A gives the proof of an intermediate technical result, Appendix B contains a tabular summary of notation for the readers convenience.

2. Proofs

Define two functionals $B, J : \mathcal{A}_n \to \mathbb{R}_0^+$ by

$$\begin{aligned} B(f) &= \max_{k\in\{1,\ldots,n\}} \sup_{\mathbf{x}\in\mathbb{B}^n, y, y'\in\mathbb{B}} D^k_{y,y'} f(\mathbf{x}) \\ J(f) &= \left(\sup_{\mathbf{x}\in\mathbb{B}^n} \sum_{k,l:k\neq l} \sup_{z,z'\in\mathbb{B}} \sup_{\mathbf{x}\in\mathbb{B}^n, y, y'\in\mathbb{B}} \left(D^l_{z,z'} D^k_{y,y'} f(\mathbf{x}) \right)^2 \right)^{1/2}. \end{aligned}$$

Observe that both B and J are positive homogeneous of order one and satisfy a triangle inequality. The fact that the second order differences in J are for distinct indices will be the key to the proof of Theorem 2. We have the following concentration results.

Theorem 4 *Let $f \in \mathcal{A}_n$, $t > 0$. Then (i)*

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-t^2}{2\sup_{\mathbf{x}\in\mathbb{B}^n} \Sigma^2(f)(\mathbf{x}) + 2B(f)\, t/3} \right)$$

and (ii)

$$\Pr\{f - Ef > t\} \leq \exp\left(\frac{-t^2}{2\sigma^2(f) + J^2(f)/2 + (2B(f)/3 + J(f))\, t} \right).$$

Part (i) appears in McDiarmid (1998) Theorem 3.8. Part (ii) is a very recent concentration inequality, Corollary 5 in Maurer (2017).

Setting the probability in part (ii) equal to $\delta \in (0, 1/e)$ and solving for the deviation t it follows, that with probability at least $1 - \delta$ in the draw of $\mathbf{x} \sim \mu^n$ we have

$$\begin{aligned} \Delta(\mathbf{x}) - E[\Delta] &\leq \sqrt{(2\sigma^2(\Delta) + J^2(\Delta)/2)\ln(1/\delta)} + (2B(\Delta)/3 + J(\Delta))\ln(1/\delta) \\ &\leq \sqrt{2\sigma^2(\Delta)\ln(1/\delta)} + 2(B(\Delta)/3 + J(\Delta))\ln(1/\delta), \end{aligned}$$

where we used $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ (several times) and $\sqrt{\ln(1/\delta)} \leq \ln(1/\delta)$, because $\delta \in (0,1/e)$. To prove Theorem 2 it therefore suffices to establish the two inequalities

$$\begin{aligned} E[\Delta] &\leq \alpha_1(\ell,\lambda)/n \quad (3) \\ 2(B(\Delta)/3 + J(\Delta)) &\leq \alpha_2(\ell,\lambda)/n. \quad (4) \end{aligned}$$

Both inequalities will follow from bounds on difference operators, and we now explain our (potentially weak) strategy to bound expressions of the form $D^k_{y,y'}f$ and $D^l_{z,z'}D^k_{y,y'}f$ for $f \in \mathcal{A}_n$.

Let B_1 and B_2 be Banach spaces, $F : B_1 \to B_2$. We write $F \in C^n$ if F is n-times Fréchet differentiable at every point of B_1. Recall that the Fréchet derivative of a function $F : B_1 \to B_2$, if it exists, at a point $x \in B_1$ is a linear map $F'(x) : B_1 \to B_2$ and that the directional derivative $F'(x)[v]$ for $v \in B_1$ can be computed by the formula $F'(x)[v] = \psi'(0)$, where $\psi_{F,x,v} : \mathbb{R} \to B_2$ is defined by $\psi_{F,x,v}(t) = F(x+tv)$. The second Fréchet derivative of F at x, if it exists, is a bilinear map $F''(x) : B_1 \times B_1 \to B_2$ and for $v, v' \in B$ we have $F''(x)[v,v'] = \partial_1\partial_2\phi_{F,x,v,v'}(0,0)$, where $\phi_{F,x,v,v'} : \mathbb{R}^2 \to B_2$ is defined by $\psi_{F,x,v,v'}(t,s) = F(x+sv+tv')$.

Now let $f \in \mathcal{A}_n \cap C^1$. For $k \in \{1,...,n\}$ and $y \in H$ use $\hat{k}(y)$ to denote the corresponding member of H^n all of whose components are zero, except the k-th one, which is equal to y. The map $\hat{k}$ is a linear isometric embedding. Then for $y, y' \in \mathbb{B}$, by the fundamental theorem of calculus

$$\begin{aligned} D^k_{y,y'}f(\mathbf{x}) &= f\left(\mathbf{x}+\hat{k}(y-x_k)\right) - f\left(\mathbf{x}+\hat{k}(y'-x_k)\right) \\ &= \int_0^1 \psi'_{f,\mathbf{x},\hat{k}(y-y')}(t)\,dt \leq \sup_{\mathbf{x}\in\mathbb{B}} \left| f'(\mathbf{x})\left[\hat{k}(y-y')\right]\right|. \end{aligned}$$

Also, if $f \in \mathcal{A}_n \cap C^2$, then for $y, y', z, z' \in \mathbb{B}$ and $l, k \in \{1,...,n\}$, likewise by the fundamental theorem of calculus

$$\begin{aligned} & D^l_{z,z'}D^k_{y,y'}f(\mathbf{x}) \\ = \; & f\left(\mathbf{x}+\hat{k}(y-x_k)+\hat{l}(z-x_l)\right) - f\left(\mathbf{x}+\hat{k}(y'-x_k)+\hat{l}(z-x_l)\right) \\ & -f\left(\mathbf{x}+\hat{k}(y-x_k)+\hat{l}(z'-x_l)\right) + f\left(\mathbf{x}+\hat{k}(y'-x_k)+\hat{l}(z-x_l)\right) \\ = \; & \int_0^1\int_0^1 \partial_1\partial_2\phi_{F,x,\hat{k}(y-y'),\hat{l}(z-z')}(t,s)\,dtds \leq \sup_{\mathbf{x}\in\mathbb{B}} \left| f''(\mathbf{x})\left[\hat{k}(y-y'),\hat{l}(z-z')\right]\right|. \end{aligned}$$

The great weakness of this method is that we uniformly bound the derivatives, while their contribution to the total difference may only be on a set of very small measure, or there may be cancellations. In Bousquet and Elisseeff (2002) for $f = \Delta$ first order differences are bounded by the methods of convex analysis instead. Since we don't know how to extend these methods to second order differences, we proceed along the more awkward path of differentiation.

Proposition 5 *Let $k, l \in \{1,...,n\}$, $k \neq l$, $y, y' \in H$, and $\|y\|, \|y'\| \leq 2$. Then for every $\mathbf{x} \in H^n$*

$$\begin{aligned} \left|L'(\mathbf{x})\left[\hat{k}(y)\right]\right| &\leq \frac{\|y\|\max\{1,c',c''\}^2}{\lambda^{3/2}n} \\ \left|\Delta'(\mathbf{x})\left[\hat{k}(y)\right]\right| &\leq \frac{3\|y\|\max\{1,c',c''\}^2}{\lambda^{3/2}n} \\ \left|\Delta''(\mathbf{x})\left[\hat{k}(y),\hat{l}(y')\right]\right| &\leq \frac{32\max\{1,c',c''\}^3}{\lambda^3 n^2} + \frac{12c'''\max\{1,c',c''\}^3}{\lambda^4 n^2} \end{aligned}$$

The proof of this proposition, which involves some tedious computations, will be given in Appendix A. Here we just show how to obtain the right order in n for the derivatives of the function g as in (1) in the simpler case of the square loss. Then we have an explicit solution of (1),

$$g(\mathbf{x}) = T^{-1}(\mathbf{x})\, z(\mathbf{x}),$$

where $T(\mathbf{x})$ is the operator $v \mapsto (1/n)\sum_{i=1}^{n} \langle v, x_i\rangle x_i + 2\lambda v$ and $z(\mathbf{x}) = (1/n)\sum_{i=1}^{n} x_i$. Now we define $\hat{g}(s,t) := g\left(\mathbf{x} + t\hat{k}(y) + s\hat{l}(y')\right)$, $\hat{T}(s,t) := T\left(\mathbf{x} + t\hat{k}(y) + s\hat{l}(y')\right)$ and $\hat{z}(s,t) := z\left(\mathbf{x} + t\hat{k}(y) + s\hat{l}(y')\right)$. Then $\hat{g} = \hat{T}^{-1}\hat{z}$. A standard argument shows that $\|\hat{g}\| \leq \lambda^{-1/2}$. Also $\hat{T}$ is invertible with $\left\|\hat{T}^{-1}\right\| \leq 1/(2\lambda)$, and $\left\|\partial_1\hat{T}\right\|, \left\|\partial_2\hat{T}\right\| \leq 4/n$, and $\|\partial_1\hat{z}\|, \|\partial_2\hat{z}\| \leq 2/n$, since $\|y\| \leq 2$, and only either the k-th or l-th term in the respective sums survive differentiation. Then, always using $\|.\|$ both for vector and operator norms,

$$\begin{aligned}\|\partial_1\hat{g}\| &= \left\|\hat{T}^{-1}\left(\partial_1\hat{T}\right)\hat{T}^{-1}\hat{z} + \hat{T}^{-1}\partial_1\hat{z}\right\| \leq \left\|\hat{T}^{-1}\right\|\left\|\partial_1\hat{T}\right\|\|\hat{g}\| + \left\|\hat{T}^{-1}\right\|\|\partial_1\hat{z}\| \\ &\leq 2/\left(\lambda^{3/2}n\right) + 1/(\lambda n) \leq 3/\left(\lambda^{3/2}n\right).\end{aligned}$$

Now $\left\|\partial_2\hat{T}^{-1}\right\| = \left\|-\hat{T}^{-1}\left(\partial_2\hat{T}\right)\hat{T}^{-1}\right\| \leq 1/\left(\lambda^2 n\right)$, and $\partial_2\partial_1\hat{T} = 0$ and $\partial_2\partial_1\hat{z} = 0$, since $k \neq l$. Thus

$$\begin{aligned}\|\partial_2\partial_1\hat{g}\| &= \left\|\left(\partial_2\hat{T}^{-1}\right)\left(\partial_1\hat{T}\right)\hat{g} + \hat{T}^{-1}\left(\partial_1\hat{T}\right)\left(\partial_1\left(\hat{T}^{-1}\hat{z}\right)\right) + \left(\partial_2\hat{T}^{-1}\right)(\partial_1\hat{z})\right\| \\ &\leq \left\|\partial_2\hat{T}^{-1}\right\|\left\|\partial_1\hat{T}\right\|\|\hat{g}\| + \left\|\hat{T}^{-1}\right\|\left\|\partial_1\hat{T}\right\|\|\partial_1\hat{g}\| + \left\|\partial_2\hat{T}^{-1}\right\|\|\partial_1\hat{z}\| \\ &\leq \frac{2}{\lambda^{3/2}n^2} + \frac{6}{\lambda^{5/2}n^2} + \frac{2}{\lambda^2 n^2} \leq \frac{10}{\lambda^{5/2}n^2}.\end{aligned}$$

By definition of $\hat{g}$ this gives the claimed order in n of the derivatives of g. Bounds on the derivatives of L and $\hat{L}$ then follow from standard differentiation rules. Appendix A gives a more general and detailed version of these arguments.

Now we use Proposition 5 to prove the inequalities (3) and (4), and thus Theorem 2. Since $\mathbb{B}$ has diameter 2 we get

$$\begin{aligned}B(\Delta) &= \max_k \sup_{\mathbf{x}\in\mathbb{B}^n} \sup_{y,y'\in\mathbb{B}} D^k_{y,y'}\Delta(\mathbf{x}) \leq \max_k \sup_{\mathbf{x}\in\mathbb{B}^n} \sup_{y,y'\in\mathbb{B}} \left|\Delta'(\mathbf{x})\left[\hat{k}(y-y')\right]\right| \\ &\leq \frac{6\max\{1,c',c''\}^2}{\lambda^{3/2}n} = \frac{3\alpha_1(\ell,\lambda)}{n}. \\ J(\Delta) &= \sqrt{n(n-1)} \max_{k\neq l} \sup_{\mathbf{x}\in\mathbb{B}^n} \sup_{z,z'\in\mathbb{B}} \sup_{y,y'\in\mathbb{B}} D^l_{z,z'} D^k_{y,y'}\Delta(\mathbf{x}) \\ &\leq n \max_{k\neq l} \sup_{\mathbf{x}\in\mathbb{B}^n} \sup_{z,z'\in\mathbb{B}} \sup_{y,y'\in\mathbb{B}} \left|\Delta''(\mathbf{x})\left[\hat{k}(y-y'), \hat{k}(z-z')\right]\right| \\ &\leq \frac{32\max\{1,c,c''\}^3}{\lambda^3 n} + \frac{12c'''\max\{1,c',c''\}^3}{\lambda^4 n},\end{aligned}$$

so

$$\begin{aligned} 2\left(B\left(\Delta\right)/3+J\left(\Delta\right)\right) &\leq \frac{68\max\left\{1,c,c''\right\}^3}{\lambda^3 n}+\frac{24c'''\max\left\{1,c',c''\right\}^3}{\lambda^4 n} \\ &= \frac{\alpha_2\left(\ell,\lambda\right)}{n}, \end{aligned}$$

which proves (4).

Also, for any given $x\in\mathbb{B}$ we have

$$\ell\left(\left\langle x,g\left(\mathbf{x}\right)\right\rangle\right)-\ell\left(\left\langle x,g\left(x_1,...,x_{k-1},x,x_{k+1},...,x_n\right)\right\rangle\right)=D^k_{x_k,x}\ell\left(\left\langle x,g\left(\mathbf{x}\right)\right\rangle\right).$$

But the expectation of the second term on the left, as $x\sim\mu$ and $\mathbf{x}\sim\mu^n$, is equal to $E_{\mathbf{x}\sim\mu^n}\left[\ell\left(\left\langle x_i,g\left(\mathbf{x}\right)\right\rangle\right)\right]$, so

$$\begin{aligned} E\left[\Delta\right] &= \frac{1}{n}\sum_k E\left[\ell\left(\left\langle x,g\left(\mathbf{x}\right)\right\rangle\right)-\ell\left(\left\langle x_k,g\left(\mathbf{x}\right)\right\rangle\right)\right]=\frac{1}{n}\sum_k E_{x\sim\mu,\mathbf{x}\sim\mu^n}\left[D^k_{x_k,x}\ell\left(\left\langle x,g\left(\mathbf{x}\right)\right\rangle\right)\right] \\ &\leq \frac{1}{n}\sum_k \sup_{\mathbf{x}\in\mathbb{B}^n}\sup_{y,y'\in\mathbb{B}} E_{x\sim\mu}\left[D^k_{y,y'}\ell\left(\left\langle x,g\left(\mathbf{x}\right)\right\rangle\right)\right]=\max_k\sup_{\mathbf{x}\in\mathbb{B}^n}\sup_{y,y'\in\mathbb{B}} D^k_{y,y'}L\left(\mathbf{x}\right) \\ &\leq \max_k\sup_{\mathbf{x}\in\mathbb{B}^n}\sup_{y,y'\in\mathbb{B}}\left|L'\left(\mathbf{x}\right)\left[\hat{k}\left(y-y'\right)\right]\right|\leq\frac{2\max\left\{1,c',c''\right\}^2}{\lambda^{3/2}n}=\frac{\alpha_1\left(\ell,\lambda\right)}{n}, \end{aligned}$$

which proves (3). This completes the proof of Theorem 2.

Proof [of Theorem 3] Substitution of the above bound on $B\left(\Delta\right)$ into the first concentration inequality of Theorem 4, solving for the deviation and using the bound (3) on $E\left[\Delta\right]$ gives the first inequality. For any $\mathbf{x}\in\mathbb{B}^n$ we have

$$\begin{aligned} \Sigma^2\left(\Delta\right)\left(\mathbf{x}\right) &= \frac{1}{2}\sum_k E_{(y,y')\sim\mu^2}\left[\left(D^k_{y,y'}\Delta\left(\mathbf{x}\right)\right)^2\right] \\ &\leq \frac{1}{2}\sum_k E_{(y,y')\sim\mu^2}\left[\sup_{\mathbf{x}\in\mathbb{B}^n}\left|\Delta'\left(\mathbf{x}\right)\hat{k}\left(y-y'\right)\right|^2\right] \\ &\leq \frac{1}{2}\sum_k E_{(y,y')\sim\mu^2}\left[\left(\frac{6\max\left\{1,c',c''\right\}^2}{\lambda^{3/2}n}\right)^2\left\|y-y'\right\|^2\right] \\ &= \frac{9\alpha_1^2\left(\ell,\lambda\right)\sigma^2_{x\sim\mu}\left(x\right)}{n}. \end{aligned}$$

Substitution in the first inequality then gives the second inequality. ■

References

M. Anthony and Peter Bartlett. *Learning in Neural Networks: Theoretical Foundations*. Cambridge University Press, 1999.

P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford University Press, 2013.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428, 2002.

B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

C. Houdré. The iterated jackknife estimate of variance. *Statistics and probability letters*, 35(2): 197–201, 1997.

A. Maurer. A Bernstein-type inequality for functions of bounded interaction. *ArXiv e-prints*, January 2017.

C. McDiarmid. Concentration. In *Probabilistic Methods of Algorithmic Discrete Mathematics*, pages 195–248, Berlin, 1998. Springer.

T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9): 1481–1497, 1990.

W. Rudin. *Principles of mathematical analysis*. McGraw-Hill New York, 1964.

J. M. Steele. An efron-stein inequality for nonsymmetric statistics. *The Annals of Statistics*, pages 753–758, 1986.

Appendix A. Differentiation

We will bound the derivatives of L and Δ by first bounding the derivatives of the transformation $g : \mathbb{B}^n \to H$ and then applying simple differentiation rules.

We begin with some generalities on differentiation and implicit differentiation. If Y, X, Z are Banach spaces and $g : Y \to X$ and $f : X \to Z$ are differentiable (we drop the "Fréchet") at $x \in Y$ and $g(x) \in X$ respectively, the $f \circ g$ is differentiable at x and the derivative $(f \circ g)'(x)$ can be computed by the chain rule as

$$(f \circ g)'(x)[h] = f'(g(x))\left[g'(x)[h]\right] \text{ for } h \in Y.$$

If f and g are also twice differentiable then so is $f \circ g$ and

$$(f \circ g)'' (x) \left[h, h'\right] = f'' (g(x)) \left[g'(x)[h], g'(x)\left[h'\right]\right] + f'(g(x)) \left[g''(x)\left[h, h'\right]\right]. \quad (5)$$

Now suppose that $f : Y \times X \to Y$, $f \in C^1$. Then for every $(y, x) \in Y \times X$ there are linear maps $A_1(y, x) : Y \to Y$ and $A_2(y, x) : X \to Y$, such that $f'(y, x)[(k, h)] = A_1(y, x)[k] + A_2(y, x)[h]$. If also $f \in C^2$ then the second derivative of f at (y, x) is a bilinear map $f''(y, x) : (Y \times X) \times (Y \times X) \to Y$ which can be written as a matrix (omitting the arguments (y, x))

$$f'' = \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix},$$

where $b_{1,1} : Y \times Y \to Y$, $b_{1,2} : Y \times X \to Y$, $b_{2,1} : X \times Y \to Y$ and $b_{22} : X \times X \to Y$, are all bilinear. For (k, h) and (k', h') in $Y \times X$

$$f''\left[(k, h), (k', h')\right] = b_{1,1}\left[k, k'\right] + b_{1,2}\left[k, h'\right] + b_{2,1}\left[h, k'\right] + b_{2,2}\left[k, k'\right].$$

We now summarize some properties of implicit differentiation. Most of it can be found in the literature (Theorem 9.28 in Rudin (1964)), but we reproduce it here for completeness.

Theorem 6 *Suppose that $f : Y \times X \to Y$, $f \in C^2$, that there is a function $g : X \to Y$ such that $f(g(x), x) = 0$ for all $x \in X$, and that $A_1(y, x)$ is invertible for every $(y, x) \in Y \times X$, with $\left\|A_1^{-1}(y, x)\right\|_\infty \leq K < \infty$.*

Then $g \in C^2$ and we have for every $x \in X$ with $z := (g(x), x)$ the formulas

$$g'(x)[h] = -A_1(z)^{-1} A_2(z)[h] \text{ for } h \in X \quad (6)$$

and

$$\begin{aligned} g''(x)\left[h, h'\right] &= -A_1^{-1}(z)\left(b_{1,1}(z)\left[g'(x)[h], g'(x)\left[h'\right]\right] - b_{1,2}(z)\left[g'(x)[h], h'\right]\right. \\ &\quad \left. -b_{2,1}(z)\left[h, g'(x)\left[h'\right]\right] + b_{2,2}(z)\left[h, h'\right]\right), \text{ for } h, h' \in X. \end{aligned} \quad (7)$$

Proof Define $F(y, x) = (f(y, x), x)$. As in the proof of the implicit function theorem (Theorem 9.28 in Rudin (1964)) one shows that F is a diffeomorphism. Since $F \in C^2$ and inverses of diffeomorphisms inherit differentiability properties $G := F^{-1} \in C^2$. Then $(g(x), x) = G(0, x)$, so $g \in C^2$.

Let $\Phi : X \to Y \times X$ be defined by $\Phi(x) := (g(x), x)$. Then for every $x \in X$ the linear operator $\Phi'(x) : X \to Y \times X$ is given by

$$\Phi'(x)[h] = \left(g'(x)[h], h\right) \text{ for } h \in X.$$

By definition of Φ we have $f \circ \Phi = 0$, so also $(f \circ \Phi)' = (f \circ \Phi)'' = 0$, so by the chain rule for $h \in X$

$$0 = (f \circ \Phi)'(x)[h] = A_1(g(x), x)\left[g'(x)[h]\right] + A_2(g(x), x)[h].$$

Formula (6) now follows from applying A_1^{-1}.

Also $\Phi''(x)$ is a bilinear map $\Phi''(x): X\times X \to Y\times X$ given by $\Phi''(x)[h,h'] = (g''(x)[h,h'],0)$ for $h, h' \in X$. Since $(f\circ\Phi)'' = 0$ we have by the second order chain rule (5)

$$\begin{aligned} 0 &= (f\circ\Phi)''(x)[h,h'] \\ &= f''(\Phi(x))\left[\Phi'(x)[h],\Phi'(x)\left[h'\right]\right] + f'(\Phi(x))\left[\Phi''(x)[h,h']\right] \\ &= f''(\Phi(x))\left[\left(g'(x)[h],h\right),\left(g'(x)\left[h'\right],h'\right)\right] + A_1(\Phi(x))\left[g''(x)[h,h']\right], \end{aligned}$$

which implies that $g''(x)[h,h'] = -A_1(\Phi(x))^{-1} f''(\Phi(x))[(g'(x)[h],h),(g'(x)[h'],h')]$. But, dropping the arguments x, and $\Phi(x) = (g(x),x) = z$,

$$f''\left[\left(g'[h],h\right),\left(g'(x)\left[h'\right],h'\right)\right] = b_{1,1}\left[g'[h],g'\left[h'\right]\right] + b_{1,2}\left[g'[h],h'\right] + b_{2,1}\left[h,g'\left[h'\right]\right] + b_{2,2}[h,h'],$$

which gives (7). ■

Proposition 7 *Let g be defined by (1), $y, y' \in H$, $\|y\|, \|y'\| \leq 2$, and $k, l \in \{1,...,n\}$, $k \neq l$. Then*
(i) $\|g(\mathbf{x})\| \leq \lambda^{-1/2}$.
(ii) If $\ell \in C^2$ then $g \in C^1$ and

$$\left\|g'(\mathbf{x})\left[\hat{k}(y)\right]\right\| \leq \frac{\|y\|\max\{c',c''\}}{\lambda^{3/2}n}.$$

(iii) If $\ell \in C^3$ then $g \in C^2$ and

$$\left\|g''(\mathbf{x})\left[\hat{k}(y),\hat{l}\left(y'\right)\right]\right\| \leq \frac{6c'''\max\{1,c',c''\}^2}{\lambda^4 n^2} + \frac{8\max\{1,c',c''\}^2}{\lambda^{5/2}n^2}.$$

In the proof we will repeatedly use crude estimates of the form $a_1(a_1+a_2) \leq 2a_1\max\{a_1,a_2\} \leq 2\max\{1,a_1,a_2\}^2$, for $a_1, a_2 \geq 0$.

Proof Since $\ell(0) = 1$ we have

$$\lambda\|g(\mathbf{x})\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}\ell(\langle x_i,0\rangle) + \lambda\|0\|^2 = 1,$$

so that $\|g(\mathbf{x})\| \leq \lambda^{-1/2}$, which is (i). This in turn implies that $\left|\ell^{(i)}(\langle z,g(\mathbf{x})\rangle)\right| \leq c^{(i)}$ for all $z \in \mathbb{B}$ and $\mathbf{x} \in \mathbb{B}^n$. Define a function $f: H\times H^n \to H$ by

$$f(w,\mathbf{x}) = \frac{1}{n}\sum_i \ell'(\langle x_i,w\rangle)x_i + 2\lambda w,$$

so if $\ell \in C^{(i+1)}$ then $f \in C^{(i)}$ and $f(g(\mathbf{x}),\mathbf{x}) = 0$ by the necessary conditions for the minimum in (1). We apply Theorem 6 with $Y = H$ and $X = H^n$.

Next we find, for the first and second derivatives of the function f, the various functions A_i and b_{ij} needed to apply Theorem 6. To find A_1 and A_2 we compute for $v \in H$

$$A_1(w,\mathbf{x})[v] = \frac{d}{dt}_{(t=0)} f(w+tv,\mathbf{x})(0) = \frac{1}{n}\sum_i \ell''(\langle x_i,w\rangle)\langle x_i,v\rangle x_i + 2\lambda v.$$

Since ℓ is assumed convex and $v \mapsto \langle x_i, v\rangle x_i$ is positive semidefinite, we retain from this that $A_1(w,\mathbf{x})$ is invertible with operator norm $\left\|A_1^{-1}(w,\mathbf{x})\right\| \leq (2\lambda)^{-1}$. Since $f \in C^{(i)}$ this shows that also $g \in C^{(i)}$. Also

$$A_2(w,\mathbf{x})\left[\hat{k}(y)\right] = \frac{d}{dt}_{(t=0)} f\left(w, \mathbf{x} + t\hat{k}(y)\right)(0) = \frac{1}{n}\left[\ell''\left(\langle x_k, w\rangle\right)\langle y, w\rangle x_k + \ell'\left(\langle x_k, w\rangle\right) y\right].$$

From Theorem 6 we get

$$\begin{aligned}\left\|g'(\mathbf{x})\left[\hat{k}(y)\right]\right\| &\leq \left\|A_1^{-1}(g(\mathbf{x}),\mathbf{x})\, A_2(g(\mathbf{x}),\mathbf{x})\left[\hat{k}(y)\right]\right\| \\ &\leq (2\lambda n)^{-1}\left\|\ell''\left(\langle x_k, g(\mathbf{x})\rangle\right)\langle y, g(\mathbf{x})\rangle x_k + \ell'\left(\langle x_k, g(\mathbf{x})\rangle\right) y\right\| \\ &\leq (2\lambda n)^{-1}\left(c''\lambda^{-1/2} + c'\right)\|y\| \leq \max\left\{c', c''\right\}\lambda^{-3/2} n^{-1}\|y\|.\end{aligned}$$

This proves (ii). Evidently the analogous bound holds for $\left\|g'(\mathbf{x})\left[\hat{l}(y')\right]\right\|$.

For the second derivatives first let $v, v' \in H$ and $\mathbf{y}, \mathbf{y}' \in H^n$. Direct computation gives for any $(w,\mathbf{x}) \in H \times H^n$

$$\begin{aligned}b_{1,1}(w,\mathbf{x})\left[v, v'\right] &= \frac{1}{n}\sum_i \ell'''\left(\langle x_i, w\rangle\right)\left\langle x_i, v'\right\rangle\langle x_i, v\rangle x_i \\ b_{1,2}(w,\mathbf{x})\left[v, \mathbf{y}'\right] &= \frac{1}{n}\sum_i \Big(\ell'''\left(\langle x_i, w\rangle\right)\left\langle y_i', w\right\rangle\langle x_i, v\rangle x_i + \\ &\quad + \ell''\left(\langle x_i, w\rangle\right)\left\langle y_i', v\right\rangle x_i + \ell''\left(\langle x_i, w\rangle\right)\langle x_i, v\rangle y_i'\Big) \\ b_{2,1}(w,\mathbf{x})\left[\mathbf{y}, v'\right] &= \frac{1}{n}\sum_i \Big(\ell'''\left(\langle x_i, w\rangle\right)\langle y_i, w\rangle\left\langle x_i, v'\right\rangle x_i + \\ &\quad + \ell''\left(\langle x_i, w\rangle\right)\left\langle y_i, v'\right\rangle x_i + \ell''\left(\langle x_i, w\rangle\right)\left\langle x_i, v'\right\rangle y_i\Big) \\ b_{2,2}(w,\mathbf{x})\left[\mathbf{y}, \mathbf{y}'\right] &= \frac{1}{n}\sum_i \Big(\ell'''\left(\langle x_i, w\rangle\right)\left\langle y_i', w\right\rangle\langle y_i, w\rangle x_i + \\ &\quad + \ell''\left(\langle x_i, w\rangle\right)\langle y_i, w\rangle y_i' + \ell''\left(\langle x_i, w\rangle\right)\left\langle y_i', w\right\rangle y_i\Big).\end{aligned}$$

Substituting $g(\mathbf{x})$ for w, $\hat{k}(y)$ for $\mathbf{y}$ and $\hat{l}(y')$ for $\mathbf{y}'$ we obtain the bounds

$$\begin{aligned}\left\|b_{1,1}(g(\mathbf{x}),\mathbf{x})\left[v, v'\right]\right\| &\leq c'''\|v\|\left\|v'\right\| \\ \left\|b_{1,2}(g(\mathbf{x}),\mathbf{x})\left[v, \hat{l}(y')\right]\right\| &\leq \frac{1}{n}\left(\frac{c'''}{\lambda^{1/2}} + 2c''\right)\|v\|\left\|y'\right\| \\ \left\|b_{2,1}(g(\mathbf{x}),\mathbf{x})\left[\hat{k}(y), v'\right]\right\| &\leq \frac{1}{n}\left(\frac{c'''}{\lambda^{1/2}} + 2c''\right)\left\|v'\right\|\|y\| \\ \left\|b_{2,2}(g(\mathbf{x}),\mathbf{x})\left[\hat{k}(y), \hat{l}(y')\right]\right\| &= 0.\end{aligned}$$

The last identity depends crucially on the assumption that $k \neq l$. Then, using (7) in Theorem 6 and substitution of the bounds in (ii) we get

$$\begin{aligned}
&\left\| g'' \left[\hat{k}(y), \hat{l}(y') \right] \right\| \\
&\leq (2\lambda)^{-1} \left(\left\| b_{1,1} \left[g' \left[\hat{k}(y) \right], g' \left[\hat{l}(y') \right] \right] \right\| + \left\| b_{1,2} \left[g' \left[\hat{k}(y) \right], \mathbf{y}' \right] \right\| + \left\| b_{2,1} \left[\hat{k}(y), g' \left[\hat{l}(y') \right] \right] \right\| \right) \\
&\leq (2\lambda)^{-1} \left(c''' \left\| g' \left[\hat{k}(y) \right] \right\| \left\| g' \left[\hat{l}(y') \right] \right\| + \frac{1}{n} \left(\frac{c'''}{\lambda^{1/2}} + 2c'' \right) \left\| g' \left[\hat{k}(y) \right] \right\| \|y'\| + \right. \\
&\left. + \frac{1}{n} \left(\frac{c'''}{\lambda^{1/2}} + 2c'' \right) \left\| g' \left[\hat{l}(y') \right] \right\| \|y\| \right) \\
&\leq \frac{2c''' \max\{c', c''\}^2}{\lambda^4 n^2} + \frac{4c''' \max\{c', c''\}}{\lambda^3 n^2} + \frac{8c'' \max\{c', c''\}}{\lambda^{5/2} n^2} \\
&\leq \frac{6c''' \max\{1, c', c''\}^2}{\lambda^4 n^2} + \frac{8 \max\{1, c', c''\}^2}{\lambda^{5/2} n^2}
\end{aligned}$$

In the third inequality we used the assumption $\|y\|, \|y'\| \leq 2$, then we used $\lambda \leq 1$. ■

Proof [of Proposition 5] For the empirical risk we find the derivatives

$$\begin{aligned}
\hat{L}'(\mathbf{x}) \left[\hat{k}(y) \right] &= \frac{1}{n} \ell' \left(\langle x_k, g(\mathbf{x}) \rangle \right) \langle y, g(\mathbf{x}) \rangle + \frac{1}{n} \sum_i \ell' \left(\langle x_i, g(\mathbf{x}) \rangle \right) \left\langle x_i, g'(\mathbf{x}) \left[\hat{k}(y) \right] \right\rangle \\
\hat{L}''(\mathbf{x}) \left[\hat{k}(y), \hat{l}(y') \right] &= \frac{1}{n} \ell'' \left(\langle x_l, g(\mathbf{x}) \rangle \right) \langle y', g(\mathbf{x}) \rangle \left\langle x_l, g'(\mathbf{x}) \left[\hat{k}(y) \right] \right\rangle + \\
&\quad + \frac{1}{n} \ell'' \left(\langle x_k, g(\mathbf{x}) \rangle \right) \langle y, g(\mathbf{x}) \rangle \left\langle x_k, g'(\mathbf{x}) \left[\hat{l}(y') \right] \right\rangle \\
&\quad + \frac{1}{n} \sum_i \ell'' \left(\langle x_i, g(\mathbf{x}) \rangle \right) \left\langle x_i, g'(\mathbf{x}) \left[\hat{l}(y') \right] \right\rangle \left\langle x_i, g'(\mathbf{x}) \left[\hat{k}(y) \right] \right\rangle + \\
&\quad + \frac{1}{n} \ell' \left(\langle x_k, g(\mathbf{x}) \rangle \right) \left\langle y, g'(\mathbf{x}) \left[\hat{l}(y') \right] \right\rangle + \\
&\quad + \frac{1}{n} \ell' \left(\langle x_l, g(\mathbf{x}) \rangle \right) \left\langle y', g'(\mathbf{x}) \left[\hat{k}(y) \right] \right\rangle + \\
&\quad + \frac{1}{n} \sum_i \ell' \left(\langle x_i, g(\mathbf{x}) \rangle \right) \left\langle x_i, g''(\mathbf{x}) \left[\hat{k}(y), \hat{l}(y') \right] \right\rangle
\end{aligned}$$

Substitution of the bounds on $\|g(\mathbf{x})\|$, $\left\| g'(\mathbf{x}) \left[\hat{k}(y) \right] \right\|$, $\left\| g'(\mathbf{x}) \left[\hat{l}(y') \right] \right\|$, $\|y\|$ and $\|y'\|$ gives

$$\begin{aligned}
\left| \hat{L}'(\mathbf{x}) \left[\hat{k}(y) \right] \right| &\leq \frac{2 \|y\| \max\{1, c', c''\}^2}{\lambda^{3/2} n} \\
\left| \hat{L}''(\mathbf{x}) \left[\hat{k}(y), \mathbf{y}' \right] \right| &\leq \frac{20 \max\{1, c', c''\}^3}{\lambda^3 n^2} + \frac{6c''' \max\{1, c', c''\}^3}{\lambda^4 n^2}.
\end{aligned}$$

As for the expected risk

$$\begin{aligned} L'(\mathbf{x})\left[\hat{k}(y)\right] &= E_x\left[\ell'(\langle x, g(\mathbf{x})\rangle)\left\langle x, g'(\mathbf{x})\left[\hat{k}(y)\right]\right\rangle\right] \\ L''(\mathbf{x})\left[\hat{k}(y), \hat{l}(y')\right] &= E_x\left[\ell''(\langle x, g(\mathbf{x})\rangle)\left\langle x, g'(\mathbf{x})\left[\hat{l}(y')\right]\right\rangle\left\langle x, g'(\mathbf{x})\left[\hat{k}(y)\right]\right\rangle + \right. \\ &\quad \left. + \ell'(\langle X, g(\mathbf{x})\rangle)\left\langle x, g''(\mathbf{x})\left[\hat{k}(y), \hat{l}(y')\right]\right\rangle\right] \end{aligned}$$

and substitution gives

$$\left|L'(\mathbf{x})\left[\hat{k}(y)\right]\right| \leq \frac{\|y\| \max\{1, c', c''\}^2}{\lambda^{3/2} n},$$

which is the first inequality to prove, and

$$\left|L''(\mathbf{x})\left[\hat{k}(y), \hat{l}(y')\right]\right| \leq \frac{12 \max\{1, c', c''\}^3}{\lambda^3 n^2} + \frac{6c''' \max\{1, c', c''\}^3}{\lambda^4 n^2}.$$

Combining the inequalities for L and $\hat{L}$ the proposition follows from $|\Delta'(\mathbf{x})[\mathbf{y}]| \leq |L'(\mathbf{x})[\mathbf{y}]| + \left|\hat{L}'(\mathbf{x})[\mathbf{y}]\right|$ and $|\Delta''(\mathbf{x})[\mathbf{y}, \mathbf{y}']| \leq |L''(\mathbf{x})[\mathbf{y}, \mathbf{y}']| + \left|\hat{L}''(\mathbf{x})[\mathbf{y}, \mathbf{y}']\right|$. ■

We conclude with an elementary bound on the variance of bounded variables.

Lemma 8 *If a random variable X has values in $[a, b]$ then $\sigma^2(X) \leq (b-a)/4$.*

Proof Since $E[X] \in [a, b]$ we have

$$\sigma^2(X) = E[(X - EX)(X - a)] \leq E[(b - EX)(X - a)] = (b - EX)(EX - a).$$

The conclusion follows from elementary calculus. ■

Appendix B. Table of notation

Symbol	short description	page
$H, \langle ., . \rangle, \|.\|$	Hilbert space with inner product and norm	1
$\|.\|$	also used for operator norm in H	1
$\mathbb{B}$	unit ball of H	1
ℓ	loss function	1
$L, \hat{L}, \Delta$	true risk, empirical risk, generalization error	1
α_1, α_2	real functions depending on ℓ and λ	2
c', c'', c'''	bounds on derivatives of ℓ	2
$\sigma^2(f)$	Variance of f	2
$\mathcal{A}_n$	algebra of bounded functions on $\mathbb{B}^n$	3
$D^k_{y,y'}$	difference operator	3
$\sigma^2_k(f)$	k-th conditional variance of f	3
$\Sigma^2(f)$	sum of conditional variances	3
$\sigma^2_{x\sim\mu}(x)$	variance of data distribution	4
B, J	bounded difference and interaction functionals	5
$C^{(i)}(X)$	space of i times differentiable functions on X	6
$F'(x)[v]$	derivative of F at x in direction v	6
$F''(x)[v, v']$	second derivative of F at x in directions v and v'	6
$\hat{k}, \hat{l}$	embedding $H \to H^n$ at k-th (l-th) coordinate	6
A_1, A_2	first derivatives in implicit differentiation	10
$b_{1,1}, b_{1,2}, b_{2,1}, b_{2,2}$	second derivatives in implicit differentiation	10