Chapter 8: Kernel learning

1. Motivation: the Kernel trick and dimension-free generalization
2. Introduction to Hilbert spaces
3. Representer Theorem
4. Positive definite kernels and the reproducing kernel Hilbert space (RKHS)
5. Moore-Aronszajn theorem and continuity of point evaluations
6. Translation-invariant kernels on $\mathbb{R}^d$
7. Generalization properties

Motivation

Key Idea: Suppose that we observe i.i.d. samples $(a_i, b_i) \sim \mathcal{P}$

- it may be that b_i are not well-approximated by a linear function of a_i,
- but b_i may be nearly linear in $\varphi(a_i)$, where a feature map $\varphi\colon \mathbb{R}^d \to \mathcal{H}$ maps to a large (infinite) dimensional space $\mathcal{H}$.

Informally, we will then encounter problems of the form

$$\min_{f \in \mathcal{H}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(b_i, \langle f, \varphi(a_i) \rangle) + \lambda \|f\|_{\mathcal{H}}^2.$$

This seems difficult because this is an infinite-dimensional problem ...

Motivation

Problem:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(b_i, \langle f, \varphi(a_i) \rangle) + \lambda \|f\|_{\mathcal{H}}^2.$$

The are two issues to discuss:

1. **Computation:** We will see that a solution $f \in \mathcal{H}$ will lie in the span of $\{\varphi(a_i)\}_i$ and therefore the whole problem reduces to

$$\min_{y \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(b_i, \sum_{j=1}^{n} y_j \langle \varphi(a_j), \varphi(a_i) \rangle) + \lambda \sum_{i,j=1}^{n} y_i y_j \langle \varphi(a_i), \varphi(a_j) \rangle.$$

   Thus, if we can evaluate the **Kernel function** $(a_j, a_i) \mapsto K(a_j, a_i)$, the problem becomes finite dimensional ! This is called the "Kernel Trick."

2. **Generalization:** Our generalization bounds based on (1) Rademacher complexity for linear classes and (2) convexity and regularization were dimension independent and therefore apply directly.

Hilbert spaces

We will need to introduce some basic functional analysis.

Defn: An inner product $\langle \cdot, \cdot \rangle$ on a vector space $\mathcal{H}$ is a mapping from $\mathcal{H} \times \mathcal{H}$ to $\mathbb{R}$ such that for all $f, g, h \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$ the following hold:

- $\langle f, g \rangle = \langle g, f \rangle$ [Symmetry]
- $\langle f, f \rangle \geq 0$ with equality if and only if $f = 0$ [Positivity]
- $\langle \alpha f + \beta g, h \rangle = \alpha \langle f, h \rangle + \beta \langle g, h \rangle$ [Linearity]

The function $\|f\| := \sqrt{\langle f, f \rangle}$ is called the induced norm.

A sequence $\{f_i\} \subset \mathcal{H}$ is called Cauchy if for all $\epsilon > 0$ we have

$$\|f_i - f_j\| \leq \epsilon$$

for all sufficiently large i and j. The vector space $\mathcal{H}$ is called a Hilbert Space if any Cauchy sequence in $\mathcal{H}$ is guaranteed to converge to some element in $\mathcal{H}$.

Hilbert spaces

Example: Sequence space $\ell^2(\mathbb{N})$ is

$$\ell^2(\mathbb{N}) := \{(x_i)_{i=1}^{\infty} : \sum_{i=1}^{n} x_i^2 < \infty\}$$

equipped with the inner product $\langle x, y\rangle = \sum_{i=1}^{\infty} x_i y_i$.

Example: The square integrable functions

$$L^2([0,1]) := \left\{ f\colon [0,1] \to \mathbb{R} : \int_0^1 f^2(s)\, ds < \infty \right\}$$

equipped with the inner product $\langle f, g\rangle = \int_{i=0}^{1} f(s)g(s)\, ds$.

Nonexample: The subspace of $\ell^2(\mathbb{N})$ consisting of all sequences with finite support is not complete (why?).

Basic properties

Cauchy–Schwarz inequality: All $f, g \in \mathcal{H}$ satisfy

$$|\langle f, g \rangle| \leq \|f\| \cdot \|g\|$$

with equality if and only if f and g are collinear.

Pythagorean Theorem: If $S \subset \mathcal{H}$ is a finite set of pairwise orthogonal elements, then

$$\left\| \sum_{f \in S} f \right\|^2 = \sum_{f \in S} \|f\|^2.$$

Closeness: For any set $Q \subset \mathcal{H}$, the set $Q^\perp$ is a closed linear subspace.

Orthogonal Decomposition: For any closed linear subspace $\mathcal{V} \subset \mathcal{H}$ and any $f \in \mathcal{H}$, there exist unique elements $f_1 \in \mathcal{V}$ and $f_2 \in \mathcal{V}^\perp$ satisfying $f = f_1 + f_2$.

Separable Hilbert Spaces

Defn. A Hilbert space $\mathcal{H}$ is called separable if it has a countable orthonormal basis, that is there exists a countable orthonormal set of vectors $\{f_i\}_{i\in\mathbb{N}}$ such that any $f \in \mathcal{H}$ can be written as

$$f = \sum_{i=1}^{\infty} \alpha_i f_i \qquad \text{for some } \alpha \in \ell^2(\mathbb{N}).$$

Remark:

1. Both $\ell^2(\mathbb{N})$ and $L^2[0,1]$ are separable.
2. All infinite dimensional separable Hilbert spaces are isomorphic to $\ell^2(\mathbb{N})$.

Riesz representation theorem

A linear function $L\colon \mathcal{H} \to \mathbb{R}$ is called bounded if there exists $M < \infty$ such that

$$|L(f)| \leq M\|f\| \qquad \forall f \in \mathcal{H}.$$

One can show (do it!) that a linear functional is bounded iff it is continuous.

Theorem (Riesz representation theorem)

Let $L\colon \mathcal{H} \to \mathbb{R}$ be a bounded linear functional. Then there exists a unique element $g \in \mathcal{H}$ such that $L(f) = \langle g, f \rangle$ for all $f \in \mathcal{H}$.

Thus we can identify all bounded linear functionals with elements of $\mathcal{H}$.

Representer Theorem

The following theorem allows to reduce typical optimization problems over $\mathcal{H}$ to optimization over a finite dimensional subspace.

Theorem (Representer)

Fix an arbitrary set $\mathcal{X}$ and a Hilbert space $\mathcal{H}$. Let $\{a_1, \ldots, a_n\} \subset \mathcal{X}^n$ be arbitrary and let $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ be any function that is non-decreasing in the last coordinate. Then the two values

$$\inf_{f \in \mathcal{H}} \Psi(\langle f, \varphi(a_1)\rangle, \ldots, \langle f, \varphi(a_n)\rangle, \|f\|^2) \tag{27}$$

and

$$\inf_{f \in \operatorname{span}\{\varphi(a_i)\}_{i=1}^n} \Psi(\langle f, \varphi(a_1)\rangle, \ldots, \langle f, \varphi(a_n)\rangle, \|f\|^2)$$

are equal.

The typical examples to think about are Ψ being a regularized loss function (e.g. least squares or logistic).

Proof

Set $\mathcal{V} := \operatorname{span}(\varphi(a_1), \ldots, \varphi(a_n))$. Any $f \in \mathcal{H}$ may be uniquely written as $f = f_{\mathcal{V}} + f_{\perp}$ for some $f_{\mathcal{V}} \in \mathcal{V}$ and $f_{\perp} \in \mathcal{V}^{\perp}$. Thus for all i we have

$$\langle f, \varphi(a_1) \rangle = \langle f_{\mathcal{V}}, \varphi(a_i) \rangle + \underbrace{\langle f_{\perp}, \varphi(a_i) \rangle}_{=0}.$$

From Pythagorean theorem, we have $\|f\|^2 = \|f_{\mathcal{V}}\|^2 + \|f_{\perp}\|^2$. Therefore

$$\begin{aligned} &\Psi(\langle f, \varphi(a_1) \rangle, \ldots, \langle f, \varphi(a_n) \rangle, \|f\|^2) \\ &\quad \geq \Psi(\langle f_{\mathcal{V}}, \varphi(a_1) \rangle, \ldots, \langle f_{\mathcal{V}}, \varphi(a_n) \rangle, \|f_{\mathcal{V}}\|^2), \end{aligned}$$

as claimed. □

The kernel trick

In particular, if we define the Kernel matrix $K_{i,j} = \langle \varphi(a_i), \varphi(a_j) \rangle$, then the optimization problem (27) is equivalent to

$$\inf_{y \in \mathbb{R}^n} \Psi((Ky)_1, (Ky)_2, \ldots, (Ky)_2, \|y\|_K^2)$$

meaning that the optimal f^* can be constructed from the the optimal y^* as

$$f^* = \sum_{i=1}^n y_i^* \varphi(a_i),$$

Moreover, for any $a \in \mathcal{X}$ we may write the prediction function

$$\langle f^*, \varphi(a) \rangle = \sum_{i=1}^n y_i^* K(a, a_i)$$

where we define the kernel map $K(a, a') = \langle \varphi(a), \varphi(a') \rangle$. Thus optimizing over the hypothesis class $\{a \mapsto \langle f, \varphi(a) \rangle\}_{f \in \mathcal{H}}$ is the same as optimizing over all finite sums of the form

$$a \mapsto \sum_{i=1}^n y_i K(a, a_i).$$

Positive definite kernels

Often, we would like to reason in reverse. That is, we would like to fit the observed data $\{(a_i, b_i)\}_{i=1}^n$ using functions of the form $a \mapsto \sum_{i=1}^n y_i K(a, a_i)$, for some bivariate function $K(\cdot, \cdot)$. In order to speak about generalization, we would like to know which $K(\cdot, \cdot)$ can be written as $K(a, a') = \langle \varphi(a), \varphi(a') \rangle$ for some feature map $\varphi \colon \mathcal{X} \to \mathcal{H}$. The following is the key definition.

Definition (PSD kernel)

A symmetric bivariate function $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive semidefinite (PSD) kernel if for all integers $n \geq 1$ and any elements $a_1, \ldots, a_n \in \mathcal{X}$, the matrix $\{K(a_i, a_j)\}_{i,j=1}^n$ is positive semidefinite.

Example: If K has the form $K(a, a') = \langle \varphi(a), \varphi(a') \rangle$ for some feature map $\varphi \colon \mathcal{X} \to \mathcal{H}$, then K is PSD since

$$\sum_{i,j=1}^n y_i y_j K(a_i, a_j) = \sum_{i,j=1}^n y_i y_j \langle \varphi(a_i), \varphi(a_j) \rangle = \left\| \sum_{i,j=1}^n y_i \varphi(a_i) \right\|^2 \geq 0,$$

for all $y_1, \ldots, y_n \in \mathbb{R}$ and $a_1, \ldots, a_n \in \mathcal{X}$.

We will see shortly that all PSD kernels arise in this way !

Examples

Example: [Linear kernel]

$$K(a, a') = \langle a, a' \rangle$$

Example: [Homogeneous polynomial kernel] Define

$$K(a, a') = \langle a, a' \rangle^m$$

for some fixed $m \geq 2$. We may write

$$K(a, a') = \left(\sum_{i=1}^{d} a_i a_i' \right)^m = \sum_{r_1 + \ldots + r_d = m} B_{r_1, \ldots, r_d} (a_1 a_1')^{r_1} \ldots (a_d a_d')^{r_d}$$

where $B_{r_1, \ldots, r_d}$ are the binomial coefficients. So we may explicitly write $K(a, a') = \langle \varphi(a), \varphi(a') \rangle$ for the feature map

$$\varphi(a) = (\sqrt{B_{r_1, \ldots, r_d}} a_1^{r_1} \cdots a_d^{r_d})_{r_1 + \ldots + r_d = m}.$$

Reproducing kernel Hilbert space (RKHS)

A natural guess at a feature map $\varphi(\cdot)$ that may represent a PSD kernel is the function $\varphi(x) = K(\cdot, x)$. The goal is now to construct a Hilbert space $\mathcal{H}$ that contains all such functions and satisfies the key relation

$$K(x, y) = \langle K(\cdot, x), K(\cdot, y) \rangle.$$

Definition (RKHS associated to a kernel)

Let $K(\cdot, \cdot)$ be a PSD kernel on some set $\mathcal{X}$. A Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ is called a reproducing kernel Hilbert space (RKHS) associated with K if for any $x \in \mathcal{X}$ the function $K(\cdot, x)$ lies in $\mathcal{H}$ and we have

$$\langle f, K(\cdot, x) \rangle = f(x) \qquad \forall f \in \mathcal{H}.$$

In this case, if we define the feature map $\varphi(x) = K(\cdot, x)$, we have as needed:

$$\langle \varphi(y), \varphi(x) \rangle = \langle K(\cdot, y), K(\cdot, x) \rangle = K(x, y).$$

Theorem (From kernels to features (Moore-Aronszajn))

Given any PSD kernel $K(\cdot, \cdot)$, there exists a unique RKHS associated to K.

Proof

Define the set of function

$$\tilde{\mathcal{H}} = \left\{ \sum_{j=1}^{n} \alpha_j K(\cdot, x_j) : x_1, \ldots, x_n \in \mathcal{X}, \alpha \in \mathbb{R}^n, n \in \mathbb{N} \right\}.$$

We may define the inner product between $f = \sum_{j=1}^{n} \alpha_j K(\cdot, x_j)$ and $\bar{f} = \sum_{j=1}^{\bar{n}} \bar{\alpha}_j K(\cdot, \bar{x}_j)$ by the expression

$$\langle f, \bar{f} \rangle = \sum_{i,j} \alpha_i \bar{\alpha}_j K(x_i, \bar{x}_j).$$

This expression does not depend on the representation of f since $\langle f, \bar{f} \rangle = \sum_i \bar{\alpha}_j f(\bar{x}_j)$. Let us check that this is an inner product. Symmetry and linearity follow trivially.

Proof continued

The inequality $\langle f, f \rangle \geq 0$ follows from the fact that K is PSD. Suppose now $\langle f, f \rangle = 0$, that is

$$0 = \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j).$$

Take any $(\alpha_0, x) \in \mathbb{R} \times \mathcal{X}$. We have

$$0 \leq \left\| \alpha_0 K(\cdot, x) + \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) \right\|^2 = \alpha_0^2 K(x, x) + 2\alpha_0 \sum_{i=1}^{n} \alpha_i K(x, x_i).$$

Letting α_0 tend to zero, we deduce that $\sum_{i=1}^{n} \alpha_i K(x, x_i) = 0$. Thus $\langle \cdot, \cdot \rangle$ is an inner product on $\tilde{\mathcal{H}}$. The space $\tilde{\mathcal{H}}$ can be enlarged to a complete inner product space $\mathcal{H}$, and therefore a Hilbert space. I'll omit the details.

To see uniqueness, suppose that $\mathcal{G}$ is another RKHS for K. Then clearly $\mathcal{H} \subset \mathcal{G}$. Let us write $\mathcal{G} = \mathcal{H} \oplus \mathcal{H}^\perp$. Let $g \in \mathcal{H}^\perp$ be arbitrary. Then for any x we must have

$$0 = \langle g, K(\cdot, x) \rangle = g(x)$$

and therefore $g = 0$. Thus $\mathcal{H}^\perp = \{0\}$ and $\mathcal{H} = \mathcal{G}$. □

RKHS and point evaluations

The RKHS property implies that for any $x \in \mathcal{X}$, the evaluation map $E_x(f) = f(x)$ is a bounded linear functional on the RKHS $\mathcal{H}$, since it is represented by $K(\cdot, x)$. It turns out this property characterizes RKHS.

Definition (Continuity of point evaluations)

An RKHS $\mathcal{H}$ is a Hilbert space of functions on $\mathcal{X}$ such that for every $x \in \mathcal{X}$, the point evaluation $E_x : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional.

An RKHS is automatically generated by a unique kernel.

Theorem (Building a kernel for RKHS)

For any RKHS on $\mathcal{X}$, there exists a unique PSD kernel $K(\cdot, \cdot)$ that induces it.

Proof

By Riesz representation, any point evaluation E_x is represented by some $f_x \in \mathcal{H}$. Define $K(x,y) := \langle f_x, f_y \rangle$. Let us see that K is a kernel. Clearly, K is symmetric and we compute

$$\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) = \sum_{i,j} \alpha_i \alpha_j \langle f_{x_i}, f_{x_j} \rangle = \|f_{\bar{x}}\|^2 \geq 0,$$

where $\bar{x} = \sum_i \alpha_i x_i$. The RKHS property follows from the computation

$$K(x,y) = f_y(x) \qquad \text{and} \qquad \langle f, K(\cdot, y) \rangle = \langle f, f_y \rangle = f(y).$$

Uniqueness of the kernel follows quickly (check this!). □

Examples

Nonexample: [$L^2[0,1]$] The space $L^2[0,1]$ is not an RKHS on $[0,1]$ because there is no function g_x satisfying

$$\int_0^1 g_x(y)f(y) = f(x) \qquad \forall f \in L^2[0,1].$$

Sobolev space: [$H^1[0,1]$] Let $H^1[0,1]$ be the space of absolutely continuous functions $f\colon [0,1] \to \mathbb{R}$ with $f(0) = 0$, and $f' \in L^2[0,1]$ with

$$\langle f, g\rangle = \int_0^1 f'(t)g'(t)\,dt.$$

This is an RKHS. To see this, for any x define $g_x(t) = \min\{x,t\}$. Then

$$\langle f, g_x\rangle = \int_0^x f'(t)\,dt = f(x).$$

Then kernel is then $K(z,x) = \langle g_x, g_z\rangle = g_x(z) = \min\{x,z\}$.

Translation-invariant kernels on $\mathbb{R}^d$

Numerous kernels have the form $K(x, x') = q(x - x')$ where $q \colon \mathbb{R}^d \to \mathbb{R}$ is a translation-invariant function. Positive semi-definiteness of such kernels can be characterized using the Fourier transform:

$$\hat{q}(w) = \int_{\mathbb{R}^d} e^{-i\langle w, x\rangle} q(x)\, dx$$

Thm: (Böchner) If q is Lebesgue integrable and its Fourier transform only takes non-negative real values, then $K(x, x') = q(x - x')$ is a PSD kernel. Moreover, the norm in the corresponding RKHS is given by

$$\|f\|^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(w)|^2}{\hat{q}(w)}\, dw.$$

This expression has an intuitive meaning when $1/\hat{q}(w)$ is a polynomial due to Parseval's theorem:

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |w^{[r]}|^2 \cdot |\hat{f}(w)|^2\, dw = \int_{\mathbb{R}^d} \left| \frac{\partial^{[r]} f}{\partial x^{[r]}}(x) \right|^2 dx$$

where we use the multi-index notation $[r] = (r_1, r_2, \ldots, r_d)$.

Proof

The inverse Fourier transform formula gives:

$$q(x - x') = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\langle w, x - x' \rangle} \hat{q}(w)\, dw.$$

Then we compute

$$\begin{aligned}
\sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) &= \frac{1}{(2\pi)^d} \sum_{i,j=1}^{n} \alpha_i \alpha_j \int_{\mathbb{R}^d} e^{i\langle w, x_i - x_j \rangle} \hat{q}(w)\, dw \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sum_{i,j=1}^{n} \alpha_i \alpha_j e^{i\langle w, x_i \rangle} \overline{e^{i\langle w, x_j \rangle}} \hat{q}(w)\, dw \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_{i}^{n} \alpha_i e^{i\langle w, x_i \rangle} \right|^2 \hat{q}(w)\, dw \geq 0,
\end{aligned}$$

as needed.

Now consider any function of the form $f(x) = \sum_i y_i K(x, x_i)$. Then we have the identity

$$\hat{f}(w) = \sum_i y_i e^{-i\langle w, x_i \rangle} \hat{q}(w)$$

and therefore

$$\begin{aligned} \|f\|^2 = \sum_{i,j=1}^n \alpha_i \alpha_j K(x_i, x_j) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_i^n \alpha_i e^{i\langle w, x_i \rangle} \right|^2 \hat{q}(w)\, dw \\ &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(w)|^2}{\hat{q}(w)}\, dw, \end{aligned}$$

as claimed. □

Examples

Example: [Laplace Kernel] The Laplace kernel is given by

$$K(x,y) = \exp\left(-\frac{\|x-y\|}{\sigma}\right),$$

where $\sigma > 0$ is called the bandwidth. The Fourier transform of $q(z) = \exp(-\|z\|/\sigma)$ is

$$\hat{q}(w) = 2^d \pi^{\frac{d-1}{2}} \Gamma\left(\frac{d+1}{2}\right) \frac{\sigma^{-1}}{(\sigma^{-2} + \|w\|_2^2)^{(d+1)/2}}.$$

Therefore $K(\cdot)$ is a PSD kernel and the norm on RKHS penalizes all the derivatives of f of order up to $(d+1)/2$.

Example: [Gaussian Kernel] The Gaussian kernel is given by

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{\sigma^2}\right),$$

The Fourier transform of $q(z) = \exp(-\|z\|/\sigma^2)$ is

$$\hat{q}(w) = (\pi\sigma^2)^{d/2} \exp(-\sigma^2\|w\|^2/4).$$

Therefore $K(\cdot)$ is a PSD kernel. Expanding $1/\hat{q}(w)$ as a Taylor series, we see that the norm on RKHS penalizes all derivatives.

Generalization

Let us now apply what we have learned previously to obtain generalization guarantees for learning with kernels. Namely, suppose we want to solve:

$$\min_{f \in \mathcal{H}} \; L(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{P}} \ell(y, \langle f, \varphi(x) \rangle).$$

We will suppose that $\ell(y, \cdot)$ is G-Lipschitz continuous for each y and let f^* be the minimizer of L. Suppose that $\|\varphi(x)\|^2 = K(x,x) \leq R^2$ almost surely.

There are two approaches: constrained ERM

$$\min_{f \in \mathcal{H}} \; L_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle f, \varphi(x_i) \rangle) \qquad \text{subject to } \|f\| \leq D$$

and regularized ERM

$$\min_{f \in \mathcal{H}} \; L_n^r(f) := L_n(f) + \lambda \|f\|^2.$$

The parameters D, $\lambda > 0$ need to be chosen. Generalization of the constrained ERM problem can be analyzed with Rademacher bounds while regularized ERM can be understood using stability bounds when $\ell(y, \cdot)$ is convex.

Generalization of constrained ERM

Consider the constrained ERM:

$$\min_{f\in\mathcal{H}} \; L_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \langle f, \varphi(x_i)\rangle) \qquad \text{subject to } \|f\| \le D,$$

and let f_n^c be its minimizer. Using Rademacher complexity, we have already proved that

$$\mathbb{E}\left[L(f_n^c) - \min_{\|f\|\le R} L(f)\right] \le \frac{4GRD}{\sqrt{n}}.$$

Therefore we deduce

$$\mathbb{E}L(f_n^c) - L(f^*) \le \underbrace{\frac{4GRD}{\sqrt{n}}}_{\text{estimation error}} + \underbrace{\min_{\|f\|\le D} L(f) - L(f^*)}_{\text{approximation error}}.$$

Note that we may bound the second term as $L(f) - L(f^*) \le GR\|f - f^*\|$ and therefore

$$\mathbb{E}L(f_n^c) - L(f^*) \le \frac{4GRD}{\sqrt{n}} + GR(\|f^*\| - D)_+.$$

Generalization of regularized ERM

Consider the regularized ERM:

$$\min_{f \in \mathcal{H}} \; L_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle f, \varphi(x_i) \rangle) + \lambda \|f\|^2,$$

and let f_n^λ be its minimizer. Suppose moreover that $\ell(y, \cdot)$ is convex. Then the stability bounds we have derived give:

$$\mathbb{E} L(f_n^\lambda) - \min L \leq \frac{\lambda \|f^*\|^2}{2} + \frac{2G^2R^2}{4\lambda n}.$$

With the optimal choice of $\lambda = \sqrt{\frac{G^2R^2}{n\|f^*\|^2}}$ we get

$$\mathbb{E} L(f_n^\lambda) - \min L \leq \frac{2GR\|f^*\|}{\sqrt{n}},$$

which up to a constant is the same as the constrained ERM.