Lecture 10

7 Regularization

The key idea of regularization is to restrict the class of possible minimizers $\mathcal{F}$ of the empirical risk $R_{\text{emp}}[f]$ such that $\mathcal{F}$ becomes a compact set. We will assume throughout that $R_{\text{emp}}[f]$ is continuous in f.

7.1 The regularization risk functional

Adapted from [1, Chapter 4.1]

We do not specify a compact set $\mathcal{F}$ since this will lead to a constrained optimization problem which can be hard to solve numerically. Rather we add a regularization term $\Omega[f]$ to the empirical risk:

$$R_{\text{reg}}[f] = R_{\text{emp}}[f] + \lambda\Omega[f].$$

The scalar $\lambda \geq 0$ is the regularization parameter which specifies the trade-off between minimizing the training error via $R_{\text{emp}}[f]$ and the smoothness or simplicity of the resulting minimizer f, which is enforced by $\Omega[f]$, and can help f generalize to new points $x \in \mathcal{X}$.

Let $\Phi : \mathcal{X} \to \ell^2$ (or in $\mathbb{R}^m$) be a feature map and consider functions f of the form:

$$f(x) = \langle \Phi(x), \mathbf{w} \rangle = \sum_{j=1}^{\infty} w_j \phi_j(x), \quad \mathbf{w} = (w_j)_{j=1}^{\infty} \in \ell^2.$$

A common regularization term in this case is:

$$\Omega[f] = \frac{1}{2}\|\mathbf{w}\|^2,$$

which gives:

$$R_{\text{reg}}[f] = R_{\text{emp}}\left[\sum_{j=1}^{\infty} w_j \phi_j\right] + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{34}$$

If we are in the regression setting, then minimizing $\|\mathbf{w}\|^2$ gives us the smoothest (or flattest) f in the Φ coordinates.

It is often the case though that we don't construct the feature map explicitly, but rather only compute similarities through a kernel k. If k is a Mercer kernel, then it is also positive semidefinite and thus is the kernel of some RKHS $\mathcal{H}$. Given training data $\{(x_i, y_i)\}_{i\leq n}$, consider functions $f \in \mathcal{H}$ of the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i), \quad \alpha_i \in \mathbb{R}. \tag{35}$$

A natural regularization term in this case is:

$$\Omega[f] = \frac{1}{2}\|f\|_{\mathcal{H}}^2,$$

which then gives:

$$R_{\text{reg}}[f] = R_{\text{emp}}[f] + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2. \tag{36}$$

In fact (34) and (36) are equivalent! To see this, recall that since k is a Mercer kernel, we can write it as:

$$k(x, x') = \sum_{j=1}^{\infty} \lambda_j \psi_j(x)\psi_j(x'),$$

and thus we can define a feature map $\Phi : \mathcal{X} \to \ell^2$ as

$$\Phi(x) = \left(\sqrt{\lambda_j}\psi_j(x)\right)_{j=1}^{\infty}, \quad \phi_j(x) = \sqrt{\lambda_j}\psi_j(x),$$

in which it follows that $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ (inner product in ℓ^2). It thus follows that (35) can be rewritten as:

$$\begin{aligned} f(x) &= \sum_{i=1}^{n} \alpha_i k(x, x_i), \\ &= \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{\infty} \phi_j(x)\phi_j(x_i), \\ &= \sum_{j=1}^{\infty} \left(\sum_{i=1}^{n} \alpha_i \phi_j(x_i) \right) \phi_j(x), \\ &= \sum_{j=1}^{\infty} w_j \phi_j(x), \end{aligned} \tag{37}$$

where

$$w_j = \sum_{i=1}^{n} \alpha_i \phi_j(x_i).$$

Thus we just need to show that $\|\mathbf{w}\| = \|f\|_{\mathcal{H}}$. But this follows from the following calculation:

$$\begin{aligned}
\|\mathbf{w}\|^2 &= \sum_{j=1}^{\infty} |w_j|^2, \\
&= \sum_{j=1}^{\infty} \left(\sum_{i=1}^{n} \alpha_i \phi_j(x_i) \right)^2, \\
&= \sum_{j=1}^{\infty} \sum_{i,\ell=1}^{n} \alpha_i \alpha_\ell \phi_j(x_i) \phi_\ell(x_i), \\
&= \sum_{i,\ell=1}^{n} \alpha_i \alpha_\ell \sum_{j=1}^{\infty} \phi_j(x_i) \phi_j(x_\ell), \\
&= \sum_{i,\ell=1}^{n} \alpha_i \alpha_\ell k(x_i, x_\ell), \\
&= \|f\|_{\mathcal{H}}^2.
\end{aligned}$$

Figure 13 illustrates the role of regularization.

7.2 The Representer Theorem

Adapted from [1, Chapter 4.2]

Let $\{(x_i, y_i)\}_{i \leq n} \subset \mathcal{X} \times \mathcal{Y} \subset \mathcal{X} \times \mathbb{R}$ be training data and recall the empirical risk is:

$$R_{\text{emp}}[f] = \frac{1}{n} \sum_{i=1}^{n} c(x_i, y_i, f(x_i)),$$

where $c : \mathcal{X} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is a cost function. Let $\mathcal{H}$ be a RKHS and consider the following optimization problem:

$$\inf_{f \in \mathcal{H}} R_{\text{reg}}[f] = \inf_{f \in \mathcal{H}} R_{\text{emp}}[f] + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2.$$

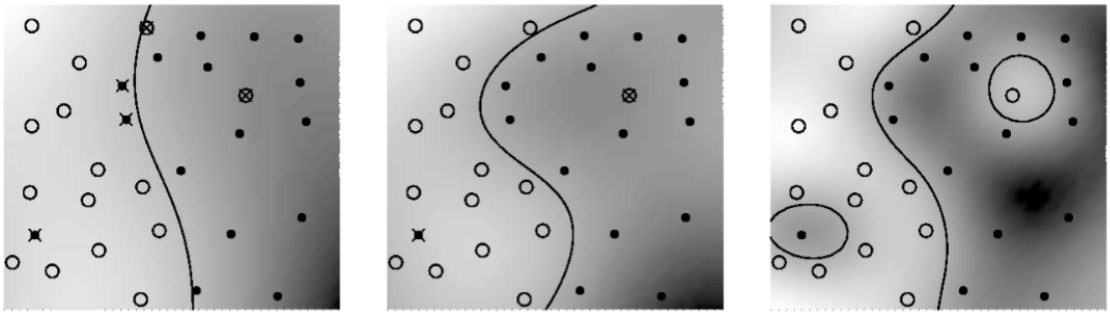

Figure 13: 2D binary classification example illustrating the role of regularization and the parameter λ. On the left, λ is very large, which yields a smooth decision boundary, but one that is perhaps too simple since it misclassifies several points. On the right, λ is small, and the decision boundary is more complex and less smooth. It correctly classifies all training points, but may not generalize well. The middle picture has a value of λ between the left and right values. The decision boundary balances smoothness and complexity, classifying most training points correctly, and will generalize well.

The solution f, a priori, can has the general form:

$$f(x) = \sum_{i=1}^{\infty} \beta_i k(x_i', x).$$

The Representer Theorem, though, guarantees that amongst all functions in our RKHS $\mathcal{H}$, the minimizer is in fact of the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x).$$

Thus even though we are trying to solve an optimization problem in a possibly infinite dimensional space, containing linear combinations of kernels centered on *arbitrary* points of $\mathcal{X}$, the solution in fact lies in the span of n particular kernels centered on the training data. We now state the theorem precisely:

Theorem 4 (The Representer Theorem)**.** *Let $\Omega : [0, \infty) \to \mathbb{R}$ be a strictly monotonic increasing function and $\mathcal{H}$ a RKHS. Then each minimizer of*

$$f^\star = \operatorname*{arg\,inf}_{f \in \mathcal{H}} \ \frac{1}{n} \sum_{i=1}^{n} c(x_i, y_i, f(x_i)) + \Omega(\|f\|_{\mathcal{H}})$$

has a representation of the form:

$$f^\star(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x). \tag{38}$$

We delay the proof till the next lecture, and give some additional remarks now. The theorem does not prevent the regularized risk from having multiple minima. To ensure a single minimum, we would need c and Ω to be convex. If Ω is merely monotonically increasing (not strictly), then it does not follow that each minimizer is of the form (38). However, one could still conclude there always exists at least one solution that does have the form (38).

References

[1] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2002.

[2] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. MIT course *Topics in Mathematics of Data Science*, 2015.

[3] Jon Shlens. A tutorial on principal component analysis. arXiv:1404.1100, 2014.

[4] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2(11):559–572, 1901.

[5] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):507–536, 1967.

[6] J. Baik, G. Ben-Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[7] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2nd edition, 2009.