High dimensional statistics for genomic data

Laurent Jacob

January 23, 2017

A little more generally : structural risk minimization

- Assume we minimize the risk over a function space $\mathcal{H}$ (polynomials of a certain degree in our example).
- If R^* is the Bayes risk, we can decompose the **Bayes regret** :

$$R(f) - R^* = \left(R(f) - \inf_{g \in \mathcal{H}} R(g) \right) + \left(\inf_{g \in \mathcal{H}} R(g) - R^* \right). \quad (1)$$

- The second term is the approximation error: the smallest excess of risk we can reach using a function of $\mathcal{H}$.
- This is a bias term, which does not depend on the data but only on the size of $\mathcal{H}$.
- The first term is the excess of risk of f with respect to the best function in $\mathcal{H}$.

A little more generally : structural risk minimization

- We consider $\hat{f}$ obtained by minimizing the empirical risk over $\mathcal{H}$:

$$\hat{f} \in \arg\min_{g \in \mathcal{H}} \hat{R}(g)$$

- We want to bound the excess of risk $R(\hat{f}) - \inf_{g \in \mathcal{H}} R(g) \geq 0$
- This term (estimation error) can be decomposed:

$$\begin{aligned} R(\hat{f}) - \inf_{g \in \mathcal{H}} R(g) \overset{\Delta}{=} & R(\hat{f}) - R(f^*_{\mathcal{H}}) \\ = & R(\hat{f}) - \hat{R}(\hat{f}) \\ & + \hat{R}(\hat{f}) - \hat{R}(f^*_{\mathcal{H}}) \\ & + \hat{R}(f^*_{\mathcal{H}}) - R(f^*_{\mathcal{H}}). \end{aligned}$$

A little more generally : structural risk minimization

$$\begin{aligned} R(\hat{f}) - \inf_{g \in \mathcal{H}} R(g) =& R(\hat{f}) - R(f^*_{\mathcal{H}}) \\ =& R(\hat{f}) - \hat{R}(\hat{f}) \\ &+ \hat{R}(\hat{f}) - \hat{R}(f^*_{\mathcal{H}}) \\ &+ \hat{R}(f^*_{\mathcal{H}}) - R(f^*_{\mathcal{H}}). \end{aligned}$$

- Reminder :
  - R is the **population** risk, $\hat{R}$ the **empirical** risk, an estimator.
  - $\hat{f}$ is the minimizer of $\hat{R}$ over $\mathcal{H}$, $f^*_{\mathcal{H}}$ is the minimizer of R over $\mathcal{H}$.
  - We therefore estimate at two levels: the function f and the risk R.

A little more generally : structural risk minimization

$$R(\hat{f}) - \inf_{g \in \mathcal{H}} R(g) = R(\hat{f}) - \hat{R}(\hat{f}) + \hat{R}(\hat{f}) - \hat{R}(f^*_{\mathcal{H}}) + \hat{R}(f^*_{\mathcal{H}}) - R(f^*_{\mathcal{H}}).$$

- The first term is the difference between the true risk and the estimated risk, for $\hat{f}$.
- This is a complex object to study. **Statistical learning theory** (Vapnik and Chervonenkis) aims at bounding this quantity as a function of n and the complexity of $\mathcal{H}$.
- The second term is nonpositive by construction.
- The third one is easier to control as it involves a deterministic function and the law of large numbers applies.

A little more generally : structural risk minimization

We can however bound the first term:

$$R(\hat{f}) - \hat{R}(\hat{f}) \leq \sup_{f \in \mathcal{H}} \left| \mathbf{E}[L(y, f(x))] - \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \right|,$$

and since this quantity also bounds the third term, we get

$$R(\hat{f}) - \inf_{g \in \mathcal{H}} R(g) \leq 2 \sup_{f \in \mathcal{H}} \left| \mathbf{E}[L(y, f(x))] - \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \right|.$$

- This bound of the estimation error suggests that it corresponds to a variance term, which increases with the size of $\mathcal{H}$.
- The more complex $\mathcal{H}$ is, the more likely it is to contain a function for which the empirical risk and the population risk are very different.

A little more generally : structural risk minimization

We can make this notion of size more precise by introducing the **Rademacher complexity** of $\mathcal{H}$:

Definition

Let ϵ_i, $i = 1, \dots, n$ i.i.d such that $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$, Z_i, $i = 1, \dots, n$ i.i.d data and $\mathcal{H}$ a space of functions defined over this data, then

$$\mathfrak{R}(\mathcal{H}) = \mathbf{E}_{\epsilon_1^n, Z_1^n}\left[\sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(Z_i) \right| \right]$$

is the Rademacher complexity of $\mathcal{H}$.

Intuition: $\mathfrak{R}$ measures the capacity of $\mathcal{H}$ to provide functions which align with noise.

A little more generally : structural risk minimization

We can make this notion of size more precise by introducing the **Rademacher complexity** of $\mathcal{H}$:

Definition

Let ϵ_i, $i = 1, \dots, n$ i.i.d such that $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$, Z_i, $i = 1, \dots, n$ i.i.d data and $\mathcal{H}$ a space of functions defined over this data, then

$$\mathfrak{R}(\mathcal{H}) = \mathbf{E}_{\epsilon_1^n, Z_1^n}\left[\sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(Z_i) \right| \right]$$

is the Rademacher complexity of $\mathcal{H}$.

This complexity increases with the size of $\mathcal{H}$ and decreases with the size n of the sample.

A little more generally : structural risk minimization

We can bound the mean estimation error in terms of the Rademacher complexity of $\mathcal{H}$.

Proposition

$$\mathbf{E}_{(x,y)_1^n} \sup_{f \in \mathcal{H}} \left| \mathbf{E}_{(x,y)}[L(y, f(x))] - \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \right| \leq 2\mathfrak{R}(\mathcal{H}).$$

Therefore,

$$\mathbf{E}_{(x,y)_1^n} \left[R(\hat{f}) - R^* \right] \leq \left(\min_{g \in \mathcal{H}} R(g) - R^* \right) + 4\mathfrak{R}(\mathcal{H}).$$

A little more generally : structural risk minimization

Therefore

$$\mathbf{E}_{(x,y)_1^n}\left[R(\hat{f}) - R^*\right] \leq \left(\min_{g\in\mathcal{H}} R(g) - R^*\right) + 4\mathfrak{R}(\mathcal{H}),$$

- This result illustrates a little more generally the bias variance tradeoff for risk minimization.
- It makes explicit the link between complexity and sample size: lots of points are needed to estimate in large $\mathcal{H}$ (otherwise $\mathfrak{R}(\mathcal{H})$ is large).

ERM consistency and SRM

Therefore

$$\mathbf{E}_{(x,y)_1^n}\left[R(\hat{f}) - R^*\right] \leq \left(\min_{g\in\mathcal{H}} R(g) - R^*\right) + 4\mathfrak{R}(\mathcal{H}),$$

Concretely, this analysis is at the core of two major elements of statistical learning (Vapnik and Chervonenkis, late 60's):

- It is used in learning theory to establish consistency of empirical risk minimization: only families with bounded complexity allow to learn by ERM (are consistent).
- **It also suggests a strategy to design estimators**: build small classes $\mathcal{H}$ which we think contain good approximations.

A little more generally : structural risk minimization

$$\mathbf{E}_{(x,y)_1^n}\left[R(\hat{f}) - R^*\right] \leq \left(\min_{g\in\mathcal{H}} R(g) - R^*\right) + 4\mathfrak{R}(\mathcal{H}),$$

Practical procedure proposed by Vapnik and Chervonenkis: **structural risk minimization**:

1. Define nested function sets of increasing complexity.
2. Minimize the empirical risk over each family.
3. Choose the solution giving the best generalization performances.

A little more generally : structural risk minimization

Structural risk minimization:

1. Define nested function sets of increasing complexity.
2. Minimize the empirical risk over each family.
3. Choose the solution giving the best generalization performances.

We will study practical instances of this strategy later in this class.

A little more generally : structural risk minimization

Proof of the previous bound (inspired from Peter Bartlett's slides)

$$\mathbf{E}_{(x,y)_1^n} \sup_{f \in \mathcal{H}} \left| \mathbf{E}_{(x,y)}[L(y, f(x))] - \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \right|$$

A little more generally : structural risk minimization

Proof of the previous bound (inspired from Peter Bartlett's slides)

$$\mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x,y)}[L(y,f(x))] - \frac{1}{n}\sum_{i=1}^{n} L(y_i,f(x_i)) \right|$$

$$= \mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x',y')_1^n} \left[\frac{1}{n}\sum_{i=1}^{n} L(y_i',f(x_i')) \right] - \frac{1}{n}\sum_{i=1}^{n} L(y_i,f(x_i)) \right|$$

A little more generally : structural risk minimization

Proof of the previous bound (inspired from Peter Bartlett's slides)

$$\mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x,y)}[L(y,f(x))] - \frac{1}{n}\sum_{i=1}^{n} L(y_i,f(x_i)) \right|$$

$$= \mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x',y')_1^n} \left[\frac{1}{n}\sum_{i=1}^{n} L(y_i',f(x_i')) \right] - \frac{1}{n}\sum_{i=1}^{n} L(y_i,f(x_i)) \right|$$

$$= \mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x',y')_1^n} \left[\frac{1}{n}\sum_{i=1}^{n} L(y_i',f(x_i')) - \frac{1}{n}\sum_{i=1}^{n} L(y_i,f(x_i)) \right] \right|$$

A little more generally : structural risk minimization

Proof of the previous bound (inspired from Peter Bartlett's slides)

$$\mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x,y)}[L(y,f(x))] - \frac{1}{n}\sum_{i=1}^n L(y_i,f(x_i)) \right|$$

$$= \mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x',y')_1^n} \left[\frac{1}{n}\sum_{i=1}^n L(y_i',f(x_i')) \right] - \frac{1}{n}\sum_{i=1}^n L(y_i,f(x_i)) \right|$$

$$= \mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x',y')_1^n} \left[\frac{1}{n}\sum_{i=1}^n L(y_i',f(x_i')) - \frac{1}{n}\sum_{i=1}^n L(y_i,f(x_i)) \right] \right|$$

$$= \mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x',y')_1^n} \left[\frac{1}{n}\sum_{i=1}^n L(y_i',f(x_i')) - L(y_i,f(x_i)) \right] \right|$$

A little more generally : structural risk minimization

Proof of the previous bound (inspired from Peter Bartlett's slides)

$$
\begin{aligned}
&\mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x,y)}[L(y,f(x))] - \frac{1}{n}\sum_{i=1}^n L(y_i,f(x_i)) \right| \\
&= \mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x',y')_1^n} \left[\frac{1}{n}\sum_{i=1}^n L(y_i',f(x_i')) \right] - \frac{1}{n}\sum_{i=1}^n L(y_i,f(x_i)) \right| \\
&= \mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x',y')_1^n} \left[\frac{1}{n}\sum_{i=1}^n L(y_i',f(x_i')) - \frac{1}{n}\sum_{i=1}^n L(y_i,f(x_i)) \right] \right| \\
&= \mathbf{E}_{(x,y)_1^n} \sup_{f\in\mathcal{H}} \left| \mathbf{E}_{(x',y')_1^n} \left[\frac{1}{n}\sum_{i=1}^n L(y_i',f(x_i')) - L(y_i,f(x_i)) \right] \right| \\
&\leq \mathbf{E}_{(x,y)_1^n} \mathbf{E}_{(x',y')_1^n} \left[\sup_{f\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^n L(y_i',f(x_i')) - L(y_i,f(x_i)) \right| \right]
\end{aligned}
$$

A little more generally : structural risk minimization

We now introduce ϵ_i, $i = 1, \dots, n \in \{-1, 1\}$. Notice that

$$\begin{aligned}
&\mathbf{E} \sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} L(y_i', f(x_i')) - L(y_i, f(x_i)) \right| \\
&= \mathbf{E} \sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left(L(y_i', f(x_i')) - L(y_i, f(x_i)) \right) \right|,
\end{aligned}$$

since the data is i.i.d, switching the two terms does not affect the distribution of the sup.
The equality holds for any choice of ϵ_i, so we can take the expectation over a uniform i.i.d choice.

A little more generally : structural risk minimization

Finally,

$$\begin{aligned}
&\mathbf{E} \sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \left(L(y_i', f(x_i')) - L(y_i, f(x_i)) \right) \right| \\
&\leq \mathbf{E} \sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i L(y_i', f(x_i')) \right| + \mathbf{E} \sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i L(y_i, f(x_i)) \right| \\
&= 2\mathbf{E} \sup_{f \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i L(y_i, f(x_i)) \right| = 2\mathfrak{R}(\mathcal{H}).
\end{aligned}$$

This proof technique is called **symmetrization**.

More intuition about the complexity of a set of functions: VC dimension

- In practice, we sometimes use VC dimension of a set of functions to bound the Rademacher complexity.
- We restrict ourselves to the sets $\mathcal{H}$ of binary valued functions (useful for classification).
- We say a set $Z = (Z_1, \ldots, Z_n)$ is **shattered** by $\mathcal{H}$ if $\mathrm{Card}\left\{f(Z_1), \ldots, f(Z_n) | f \in \mathcal{H}\right\} = 2^n$.
- Interpretation: we can find an $f \in \mathcal{H}$ assigning 0 to any subset of Z and 1 to its complement.
- The VC dimension $\nu(\mathcal{H})$ of $\mathcal{H}$ is the largest integer n such that there exists a set $(Z_1, \ldots, Z_n)$ shattered by $\mathcal{H}$.

More intuition about the complexity of a set of functions: VC dimension

- We extend the VC dimension to real valued functions by thresholding functions at 0.
- Linear functions in p dimensions: $\mathcal{H}_L = \{f_\theta(x) = sign(\theta^\top x), \theta \in \mathbb{R}^p\}$.
- Includes linear functions and polynomials in our introduction.
- We can show that $\nu(\mathcal{H}_L) = p$.

More intuition about the complexity of a set of functions: VC dimension

- Proof of $\nu(\mathcal{H}_L) \geq p$: we build a set of p points in p dimensions shattered by $\mathcal{H}_L$.

- Proof of $\nu(\mathcal{H}_L) < p+1$: no set of $p+1$ points in p dimensions can be shattered by a linear function.

More intuition about the complexity of a set of functions: VC dimension

- Proof of $\nu(\mathcal{H}_L) \geq p$: we build a set of p points in p dimensions shattered by $\mathcal{H}_L$.
  Let $\mathcal{E}_p$ be the canonical basis of $\mathbb{R}^p$. For any set $y \in \{0,1\}^p$ and any $i = 1, \ldots, n$, $f_\theta(e_i) = y_i$ by choosing $\theta_i = y_i$.
- Proof of $\nu(\mathcal{H}_L) < p + 1$: no set of $p + 1$ points in p dimensions can be shattered by a linear function.

More intuition about the complexity of a set of functions: VC dimension

- Let $x_1, \ldots, x_{p+1} \in \mathbb{R}^p$. One of the points can necessarily be written as a linear combination of the p others.

More intuition about the complexity of a set of functions: VC dimension

- Let $x_1, \dots, x_{p+1} \in \mathbb{R}^p$. One of the points can necessarily be written as a linear combination of the p others.
- Without loss of generality, let us write $x_{p+1} = \sum_{i=1}^{p} \alpha_i x_i$ and $f_\theta(x_{p+1}) = \sum_{i=1}^{p} \alpha_i \theta^\top x_i$.

More intuition about the complexity of a set of functions: VC dimension

- Let $x_1, \ldots, x_{p+1} \in \mathbb{R}^p$. One of the points can necessarily be written as a linear combination of the p others.
- Without loss of generality, let us write $x_{p+1} = \sum_{i=1}^{p} \alpha_i x_i$ and $f_\theta(x_{p+1}) = \sum_{i=1}^{p} \alpha_i \theta^\top x_i$.
- Let $y = (sign(\alpha_1), \ldots, sign(\alpha_p), -1)$, and assume there exists $\theta \in \mathbb{R}^p$ such that $sign(\theta^\top x_i) = y_i, i = 1, \ldots, p$.

More intuition about the complexity of a set of functions: VC dimension

- Let $x_1, \ldots, x_{p+1} \in \mathbb{R}^p$. One of the points can necessarily be written as a linear combination of the p others.
- Without loss of generality, let us write $x_{p+1} = \sum_{i=1}^{p} \alpha_i x_i$ and $f_\theta(x_{p+1}) = \sum_{i=1}^{p} \alpha_i \theta^\top x_i$.
- Let $y = (sign(\alpha_1), \ldots, sign(\alpha_p), -1)$, and assume there exists $\theta \in \mathbb{R}^p$ such that $sign(\theta^\top x_i) = y_i, i = 1, \ldots, p$.
- Then necessarily $sign(\theta^\top x_{p+1}) = sign(\sum_{i=1}^{p} \alpha_i \theta^\top x_i) = 1$ since $sign(\theta^\top x_i) = sign(\alpha_i), i = 1, \ldots, p$.

More intuition about the complexity of a set of functions: VC dimension

- Let $x_1, \ldots, x_{p+1} \in \mathbb{R}^p$. One of the points can necessarily be written as a linear combination of the p others.
- Without loss of generality, let us write $x_{p+1} = \sum_{i=1}^{p} \alpha_i x_i$ and $f_\theta(x_{p+1}) = \sum_{i=1}^{p} \alpha_i \theta^\top x_i$.
- Let $y = (sign(\alpha_1), \ldots, sign(\alpha_p), -1)$, and assume there exists $\theta \in \mathbb{R}^p$ such that $sign(\theta^\top x_i) = y_i, i = 1, \ldots, p$.
- Then necessarily $sign(\theta^\top x_{p+1}) = sign(\sum_{i=1}^{p} \alpha_i \theta^\top x_i) = 1$ since $sign(\theta^\top x_i) = sign(\alpha_i), i = 1, \ldots, p$.
- y can therefore not be obtained by any function of $\mathcal{H}_L$, and no set of $p+1$ vectors in $\mathbb{R}^p$ is shattered by $\mathcal{H}_L$.

Summary

- We saw how the risk could generally be decomposed as a term of bias/approximation and a term of variance/estimation.
- This decomposition highlights the tradeoff that needs to be dealt with in inference. This tradeoff is related to the complexity of the set of functions under consideration:
  - Sets too simple lead to a large approximation error.
  - Sets too large lead to a large estimation error.
- We defined this notion of complexity more precisely (Rademacher, VC), and saw it also depended on the number of samples.
- These ideas are crucial in modern applications, where we sometimes have few samples in high dimension.

Summary

- We saw how the risk could generally be decomposed as a term of bias/approximation and a term of variance/estimation.
- This decomposition highlights the tradeoff that needs to be dealt with in inference. This tradeoff is related to the complexity of the set of functions under consideration:
  - Sets too simple lead to a large approximation error.
  - Sets too large lead to a large estimation error.
- We defined this notion of complexity more precisely (Rademacher, VC), and saw it also depended on the number of samples.
- These ideas are crucial in modern applications, where we sometimes have few samples in high dimension.

Summary

- We saw how the risk could generally be decomposed as a term of bias/approximation and a term of variance/estimation.
- This decomposition highlights the tradeoff that needs to be dealt with in inference. This tradeoff is related to the complexity of the set of functions under consideration:
  - Sets too simple lead to a large approximation error.
  - Sets too large lead to a large estimation error.
- We defined this notion of complexity more precisely (Rademacher, VC), and saw it also depended on the number of samples.
- These ideas are crucial in modern applications, where we sometimes have few samples in high dimension.

Summary

- We saw how the risk could generally be decomposed as a term of bias/approximation and a term of variance/estimation.
- This decomposition highlights the tradeoff that needs to be dealt with in inference. This tradeoff is related to the complexity of the set of functions under consideration:
  - Sets too simple lead to a large approximation error.
  - Sets too large lead to a large estimation error.
- We defined this notion of complexity more precisely (Rademacher, VC), and saw it also depended on the number of samples.
- These ideas are crucial in modern applications, where we sometimes have few samples in high dimension.

Part III

Supervised learning

Perspective

- With these ideas in mind, we now turn to concrete examples of statistical learning methods.
- We focus on *penalized empirical risk minimization* techniques, which explicitly implements the bias-variance tradeoff.
- Other techniques exist (and perform sometimes very well).

Supervised learning outline

1. ℓ_2 penalties.
2. Ridge regression.
3. Fundamentals of constrained optimization.
4. Support vector machines.
5. Cross validation.

First four points are related to penalized empirical risk minimization.

Penalized empirical risk minimization

Reminder: structural risk minimization

1. **Define nested function sets of increasing complexity.**
2. Minimize the empirical risk over each family.
3. Choose the solution giving the best generalization performances.

Define a complexity measure Ω for functions, and consider the classes

$$\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \ldots,$$

where $\mathcal{H}_j = \{f, \Omega(f) \leq \mu_j\}$ and $\mu_1 < \mu_2 < \ldots$.

Reminder: structural risk minimization

Define a complexity measure Ω for functions, and consider the classes

$$\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \ldots,$$

where $\mathcal{H}_j = \{f, \Omega(f) \leq \mu_j\}$ and $\mu_1 < \mu_2 < ....$
Then (step 2) we can successively solve:

$$\min_{f \in \mathcal{H}_j} \sum_{i=1}^{n} L(y_i, f(x_i)),$$

i.e., minimize the empirical risk while restricting ourselves to sets of function of increasing complexity.
Note: this results in constrained optimization problems. Solving these problems for different loss functions and function spaces is an active research area.

Remark: equivalence with a penalized estimator

- We mostly discuss penalized methods:

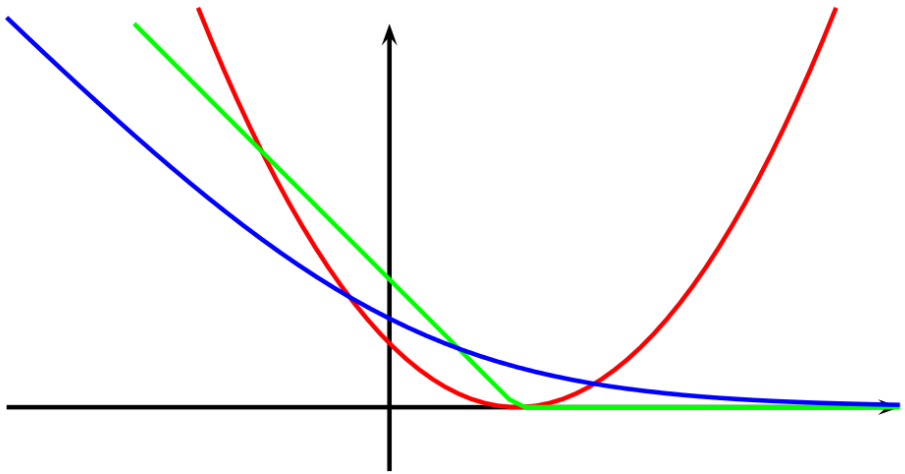
$$\min_f \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \Omega(f)$$

- The first term favors a good fit to the data, the second one favors regularity of f.
- We will show later that the constrained and penalized forms are often equivalent in some sense (need to introduce some technical tools before that).
- The approach will stay the same: we define an Ω which is relevant for our problem and we compare the generalization performances of the functions obtained for decreasing values of λ.

Usual loss functions

Regression : $y \in \mathbb{R}$

- ℓ_2 : $L(y_i, f(x_i)) = (y_i - f(x_i))^2$ (which we used in introduction),
- ℓ_1 : $L(y_i, f(x_i)) = |y_i - f(x_i)|$ (robust version, less sensitive to large errors, *e.g.* median vs mean).

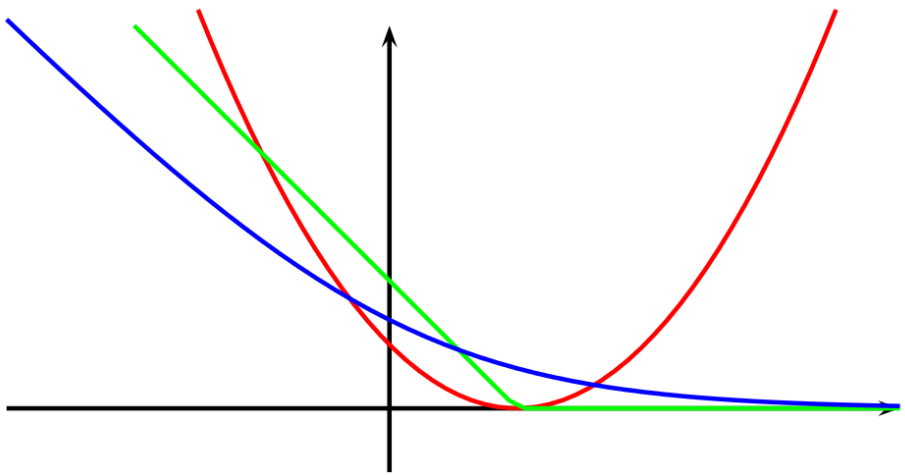


(from J. Mairal's slides)

Usual loss functions

Classification : $y \in \{0, 1\}$

- 0/1 : $L(y_i, f(x_i)) = \mathbf{1}_{y_i f(x_i) \geq 0}$,
- logistic : $L(y_i, f(x_i)) = \log\left(1 + e^{-y_i f(x_i)}\right)$,
- hinge : $L(y_i, f(x_i)) = \max(0, 1 - y_i f(x_i))$.



(from J. Mairal's slides)

Other problems: ranking, multi-class, survival...

Convexity

- 0/1 loss counts the number of missclassification, the other ones are convex approximations.
- Convex losses combined with convex penalties lead to convex objectives for which global optima can be found.
- Methods based on convex objectives are also simpler to analyze.
- However, this guarantees by no mean that the convex version of a method performs better than its non-convex counterpart in practice.

Penalties

- We now present an example of penalty, and analyze its effect on the estimated function.
- We restrict ourselves to linear functions $f(x) = \theta^\top x$, $\theta \in \mathbb{R}^p$.

Ridge penalty

- A very common penalty is the ridge :

$$\Omega(\theta) = \|\theta\|_2^2.$$

- Used in **ridge regression** combined with the ℓ_2 loss and **support vector machines** (SVM) combined with the hinge loss.
- Leads to functions with the following type of regularity: two points x, x' which are close in Euclidean norm have close evaluations by the function since by the Cauchy-Schwarz inequality,

$$|\theta^\top x - \theta^\top x'| \leq \|\theta\|_2 \|x - x'\|_2.$$

Ridge penalty

- This property can limit the overfit and improve generalization: it makes functions behave similarly over similar, potentially unobserved data.
- Of course if there is no good predictor with this kind of regularity, the risk can be high because of the approximation term.
- We now study more precisely the influence of the ridge penalty in terms of bias-variance tradeoff for the linear model:

$$y = \bar{\theta}^\top x + \varepsilon,$$

where ε is a random variable with mean zero and variance σ^2.

The ridge regression

Ridge regression: bias and variance

- We observe n realizations of the previous linear model, represented by an $X \in \mathbb{R}^{n,p}$ matrix and a $Y \in \mathbb{R}^n$ vector.
- Consider the estimator

$$\hat{\theta} = \arg\min_{\theta} \left(\|Y - X\theta\|^2 + \lambda \|\theta\|^2 \right).$$

- We can show there exists a closed form for this estimator :

$$\hat{\theta} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

- **[Exercise]** Show that the bias $\mathbf{E}[\hat{\theta} - \bar{\theta}]$ of $\hat{\theta}$ is $-\lambda(X^\top X + \lambda I)^{-1}\bar{\theta}$.

Ridge regression: bias and variance

$$
\begin{aligned}
\mathbf{E}[\hat{\theta}] &= \mathbf{E}[(X^\top X + \lambda I)^{-1} X^\top Y] \\
&= \mathbf{E}[(X^\top X + \lambda I)^{-1} X^\top \left(X\bar{\theta} + \varepsilon\right)] \\
&= (X^\top X + \lambda I)^{-1} X^\top X\bar{\theta} + (X^\top X + \lambda I)^{-1} X^\top \mathbf{E}[\varepsilon] \\
&= (X^\top X + \lambda I)^{-1} X^\top X\bar{\theta}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{E}[\hat{\theta} - \bar{\theta}] &= (X^\top X + \lambda I)^{-1} X^\top X\bar{\theta} - \bar{\theta} \\
&= \left((X^\top X + \lambda I)^{-1} X^\top X - I\right) \bar{\theta} \\
&= (X^\top X + \lambda I)^{-1} (X^\top X - X^\top X - \lambda I)\bar{\theta} \\
&= -\lambda (X^\top X + \lambda I)^{-1} \bar{\theta}.
\end{aligned}
$$

Ridge regression: bias and variance

We now look at the variance of $\hat{\theta}$:

$$\begin{aligned} Var[\hat{\theta}] &= Var[(X^\top X + \lambda I)^{-1} X^\top Y] \\ &= (X^\top X + \lambda I)^{-1} X^\top Var[Y] X (X^\top X + \lambda I)^{-1} \\ &= \sigma^2 (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1} \end{aligned}$$

(reminder : for a deterministic matrix A, $Var[AX] = AVar[X]A^\top$).

Ridge regression : bias and variance

Bias (1/3)

- The bias $-\lambda(X^\top X + \lambda I)^{-1}\bar{\theta}$ increases with λ and tends to $-\bar{\theta}$.
- Remark: $\hat{\theta} \to 0$ when $\lambda \to \infty$, so the limit bias is the one incurred by estimating $\bar{\theta}$ by 0.
- If $\lambda = 0$ (unpenalized linear regression), the bias is zero.
- The amplitude of the bias also depends on the norm of $\bar{\theta}$: if the $\bar{\theta}$ which generated the data has a small norm, the bias/approximation error incurred by restricting ourselves to small norm estimators is smaller.

Ridge regression : bias and variance

Bias (2/3)

- A little more precisely, the squared norm of the bias is ($\lambda \neq 0$) :

$$\begin{aligned} \| -\lambda(X^\top X + \lambda I)^{-1}\bar{\theta}\|^2 &= \|(\lambda^{-1}X^\top X + I)^{-1}\bar{\theta}\|^2 \\ &= \|U\Sigma U^\top \bar{\theta}\|^2 = \|\Sigma U^\top \bar{\theta}\|^2, \end{aligned}$$

  where $U\Sigma U^\top$ is the spectral decomposition of $(\lambda^{-1}X^\top X + I)^{-1}$.
- The eigenvalues of $(\lambda^{-1}X^\top X + I)^{-1}$ are **[Exercise]** :

Ridge regression : bias and variance

Bias (2/3)

- A little more precisely, the squared norm of the bias is ($\lambda \neq 0$) :

$$\begin{aligned} \| -\lambda(X^\top X + \lambda I)^{-1}\bar{\theta}\|^2 &= \|(\lambda^{-1}X^\top X + I)^{-1}\bar{\theta}\|^2 \\ &= \|U\Sigma U^\top \bar{\theta}\|^2 = \|\Sigma U^\top \bar{\theta}\|^2, \end{aligned}$$

  where $U\Sigma U^\top$ is the spectral decomposition of $(\lambda^{-1}X^\top X + I)^{-1}$.

- The eigenvalues of $(\lambda^{-1}X^\top X + I)^{-1}$ are **[Exercise]** :

$$\Sigma = \mathrm{Diag}\left(\lambda^{-1}e_i^2 + 1\right)^{-1} = \mathrm{Diag}\left(\frac{\lambda}{e_i^2 + \lambda}\right),$$

  where the e_i are the singular values of X.

Ridge regression: bias and variance

Bias (3/3)

$$\| -\lambda(X^\top X + \lambda I)^{-1}\bar{\theta}\|^2 = \left\| \mathrm{Diag}\left(\frac{\lambda}{e_i^2+\lambda}\right) U^\top \bar{\theta} \right\|^2,$$

- Provides the shape of the convergence towards maximum bias as λ increases.
- If n/p is small, $X^\top X$ typically has small eigenvalues e_i^2, and the bias is larger (even more if θ is aligned with eigenvectors corresponding to small eigenvalues).
- Statistical interpretation: the bias is larger if the vector to be estimated lies in a direction of low empirical variance of the X.