Cost Function

Recall: Simple linear regression

- Linear regression solution was defined as minimizer of $L(\boldsymbol{\beta}) := \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|^2$
- We have so far defined ridge regression only directly in terms of the estimator $\hat{\boldsymbol{\beta}}^{\text{ridge}} := (\tilde{\mathbf{X}}^t\tilde{\mathbf{X}} + \lambda\mathbb{I})^{-1}\tilde{\mathbf{X}}^t\mathbf{y}$.
- To analyze the method, it is helpful to understand it as an optimization problem.
- We ask: Which function L' does $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ minimize?

Ridge regression as an optimization problem

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\beta}\{\|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\}$$

Penalty terms

Recall: $\|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 = \sum_i L^{\text{se}}(y_i, f(\tilde{\mathbf{x}}_i; \boldsymbol{\beta}))$, so ridge regression is of the form

$$L'(\boldsymbol{\beta}) = \sum_i L^{\text{se}}(y_i, f(\tilde{\mathbf{x}}_i; \boldsymbol{\beta})) + \lambda \|\boldsymbol{\beta}\|^2$$

The term $\|\boldsymbol{\beta}\|^2$ is called a **penalty term**.

Penalized fitting

The general structure of the optimization problem is

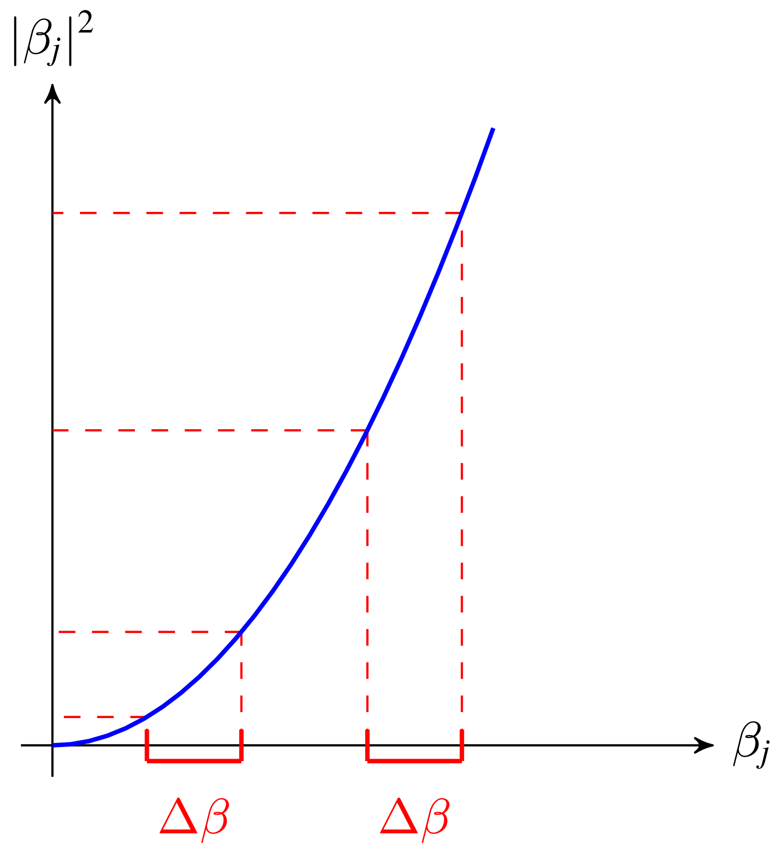
$$\text{total cost} = \text{goodness-of-fit term} + \text{penalty term}$$

Penalty terms make solutions we would like to discourage more expensive.

What kind of solutions does the choice $\|\boldsymbol{\beta}\|^2$ favor or discourage?

Quadratic Penalties

- A quadratic penalty implies that the reduction in cost we can achieve depends on the magnitude of β_j.
- Suppose we reduce β_j by a fixed amount $\Delta\beta$.
- Recall that the effect on the regression function is *linear*. The fitting cost (squared error) is quadratic, but in the *error*, not in $\boldsymbol{\beta}$.
- Consequence: Optimization algorithm will favor vectors $\boldsymbol{\beta}$ whose *entries all have similar size*.



SPARSITY

Setting

- Regression problem with n data points $\tilde{\mathbf{x}}_i$ in $\mathbb{R}^D$.
- D may be very large (much larger than n).
- Goal: Select a small subset of $d \ll D$ dimensions and discard the rest.
- In machine learning lingo: Feature selection for regression.

How do we switch off a dimension?

- In linear regression: Each entry of $\boldsymbol{\beta}$ corresponds to a dimension in data space.
- If $\beta_k = 0$, the prediction is

$$f(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \ldots + 0 \cdot x_k + \ldots + \beta_D x_D ,$$

  so the prediction does not depend on dimension k.
- Feature selection: Find a solution $\boldsymbol{\beta}$ that (1) predicts well and (2) has only a small number of non-zero entries.
- A solution in which all but a few entries vanish is called a **sparse** solution.

Penalization approach

Find a penalty term which discourages non-sparse solutions.

Can quadratic penalty help?

- Suppose β_k is large, all other β_j are small but non-zero.
- Sparsity: Penalty should keep β_k, discard others (i.e. push other β_j to zero)
- Quadratic penalty: Will favor entries β_j which all have similar size
  $\rightarrow$ pushes β_k towards small value.

Overall, a quadratic penalty favors many small, but non-zero values.

Solution

Sparsity can be achieved using *linear* penalty terms.

Sparse regression

$$\boldsymbol{\beta}^{\text{lasso}} := \arg\min_{\boldsymbol{\beta}} \{ \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \}$$

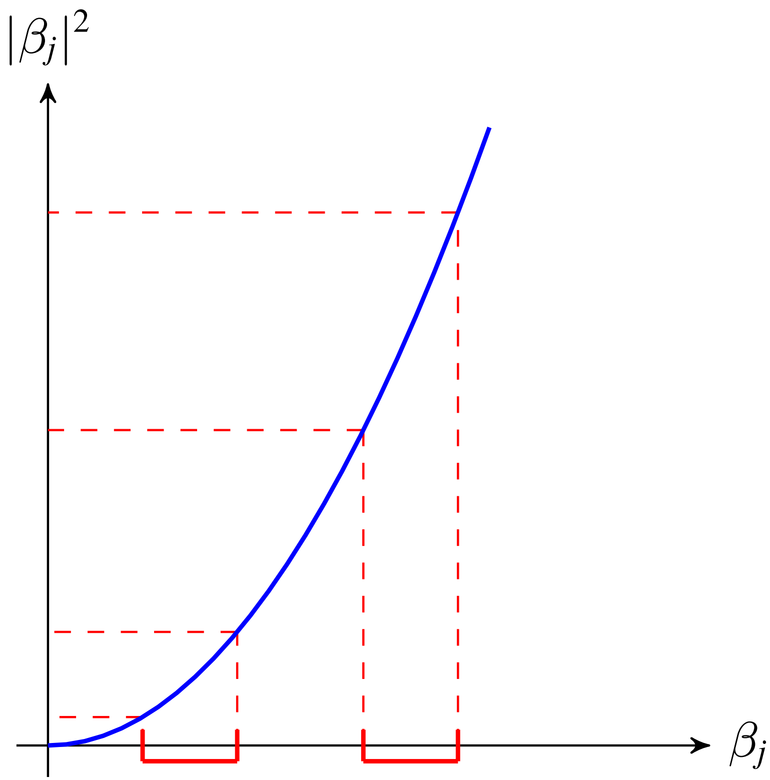
where

$$\|\boldsymbol{\beta}\|_1 := \sum_{j=1}^{D} |\beta_j|$$

The regression method which determines $\boldsymbol{\beta}^{\text{lasso}}$ is also called the LASSO (for "Least Absolute Shrinkage and Selection Operator").

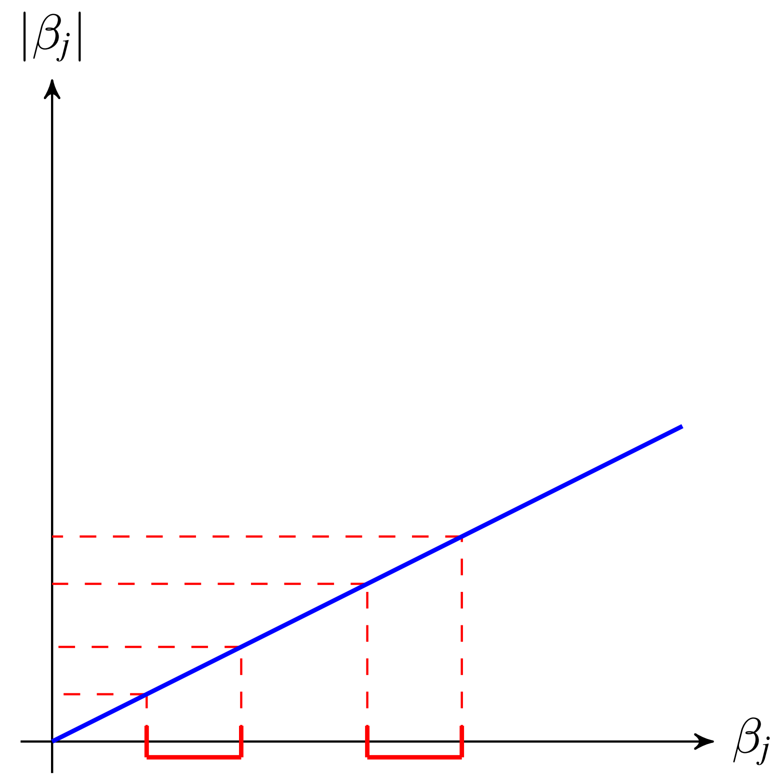
QUADRATIC PENALTIES

Quadratic penalty



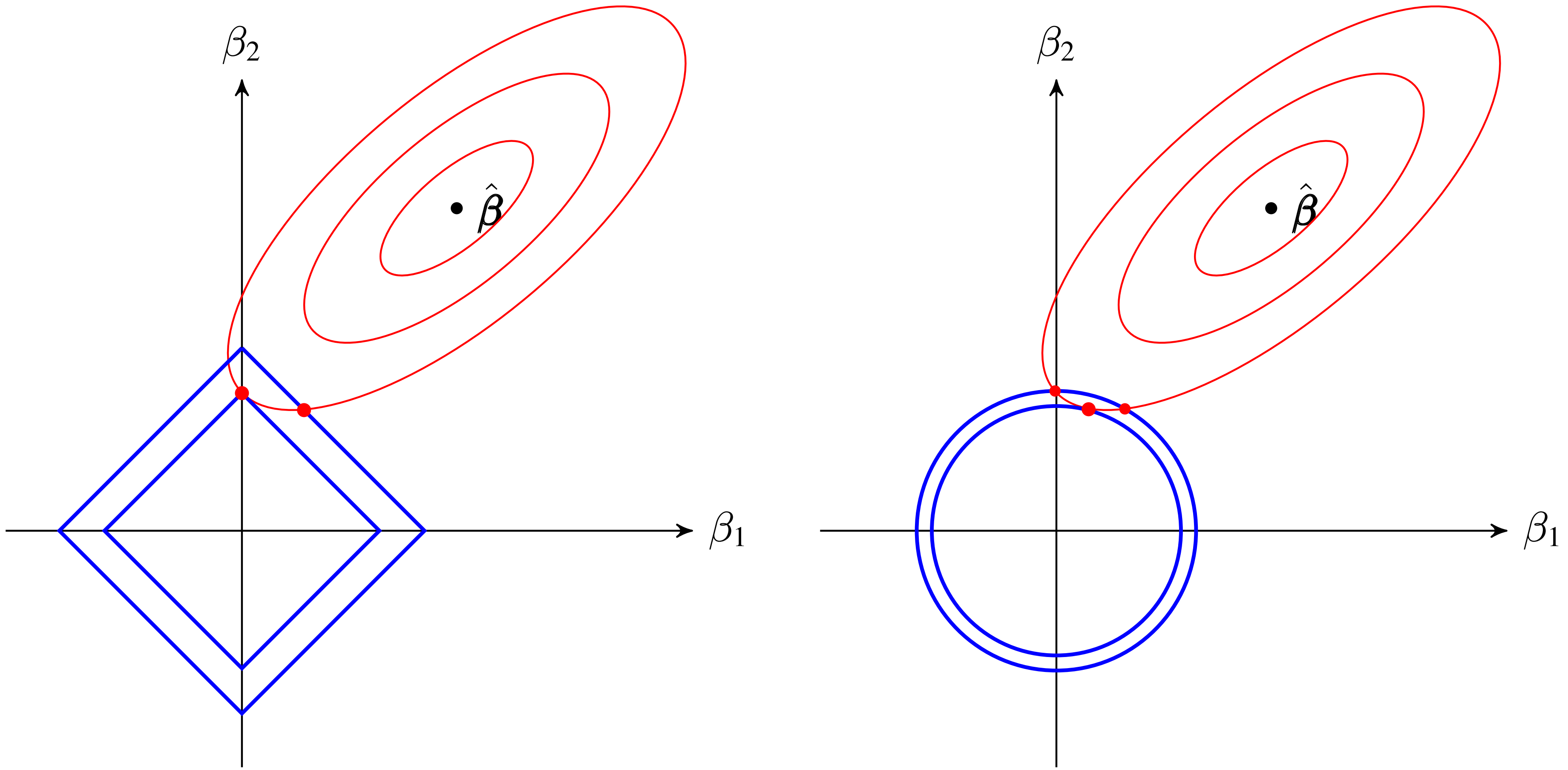
Reducing a large value β_j by a fixed amount achieves a large cost reduction.

Linear penalty



Cost reduction does not depend on the magnitude of β_j.

Ridge Regression vs Lasso



- Red: Contours of $\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2$
- Blue: Contours of $\|\boldsymbol{\beta}\|_1$ (left) and $\|\boldsymbol{\beta}\|_2$ (right)

ℓ_p Regression

ℓ_p-norms

$$\|\boldsymbol{\beta}\|_p := \Big(\sum_{j=1}^{D} |\beta_j|^p\Big)^{\frac{1}{p}} \qquad \text{for } 0 < p \leq \infty$$

is called the ℓ_p-norm.

ℓ_p-regression

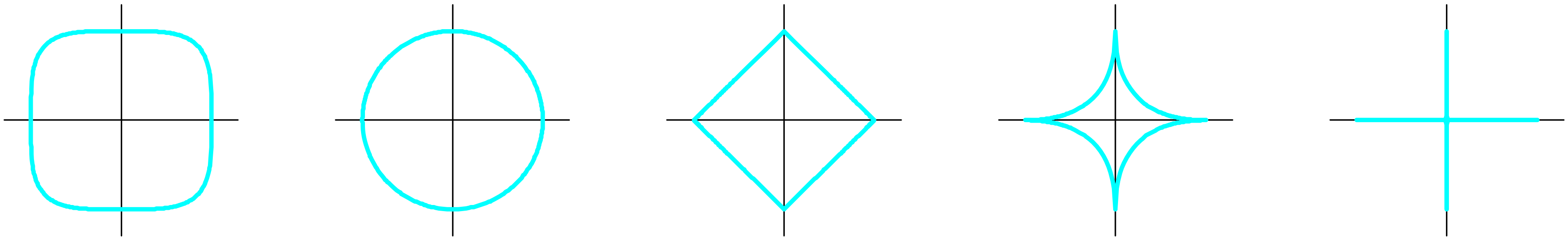
The penalized linear regression problem

$$\boldsymbol{\beta}^{\ell_p} := \arg\min_{\boldsymbol{\beta}} \{\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_p^p\}$$

is also referred to as ℓ_p**-regression**. We have seen:

- ℓ_1-regression = LASSO
- ℓ_2-regression = ridge regression

ℓ_p Penalization Terms



$p = 4$ $\quad$ $p = 2$ $\quad$ $p = 1$ $\quad$ $p = 0.5$ $\quad$ $p = 0.1$

p	Behavior of $\|\| \,.\, \|\|_p$
$p = \infty$	Norm measures largest absolute entry, $\|\|\boldsymbol{\beta}\|\|_\infty = \max_j \|\|\beta_j\|\|$
$p > 2$	Norm focusses on large entries
$p = 2$	Large entries are expensive; encourages similar-size entries.
$p = 1$	Encourages sparsity.
$p < 1$	Encourages sparsity as for $p = 1$ (note "pointy" behavior on the axes), but contour set not convex.
$p \to 0$	Simply records whether an entry is non-zero, i.e. $\|\|\boldsymbol{\beta}\|\|_0 = \sum_j \mathbb{I}\{\beta_j \neq 0\}$

Computing the Solution

Ridge regression

Recall: Solution can be computed directly as $\hat{\boldsymbol{\beta}}^{\text{ridge}} := (\tilde{\mathbf{X}}^t\tilde{\mathbf{X}} + \lambda\mathbb{I})^{-1}\tilde{\mathbf{X}}^t\mathbf{y}$. There is no similar formula for the ℓ_1 case.

Solution of ℓ_1 problem

By convex optimization.

ℓ_p Regression as an Optimization Problem

Recall: ℓ_p penalty

The optimization problem

$$\boldsymbol{\beta}^{\ell_p} := \arg\min_{\boldsymbol{\beta}} \{ \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_p^p \}$$

looks like a Lagrange version of:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 \\ \text{s.t.} \quad & \|\boldsymbol{\beta}\|_p^p \leq 0 \end{aligned}$$

However, $\|\boldsymbol{\beta}\|_p^p \leq 0$ makes no sense, since the only solution is $\boldsymbol{\beta} = (0, \dots, 0)$.

Observation

Constant shifts do not affect minima, so

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_p^p = \min_{\boldsymbol{\beta}} (\|\boldsymbol{\beta}\|_p^p - t)$$

for any $t \in \mathbb{R}$.

Formulation of Constraints

Constrained Version

$$\boldsymbol{\beta}^{\ell_p} = \arg\min_{\boldsymbol{\beta}} \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2$$

$$\text{s.t. } \|\boldsymbol{\beta}\|_p^p \leq t$$

Choosing the constraint as $\|\boldsymbol{\beta}\|_1^1 \leq t$ gives the Lasso, $\|\boldsymbol{\beta}\|_2^2 \leq t$ is ridge regression.

Feasible sets

The boundary ∂G of the feasible set is the contour set $\|\boldsymbol{\beta}\|_p^p = t$.
Recall: G is convex only if $p \geq 1$.

Summary: Regression

Methods we have discussed

- Linear regression with least squares
- Ridge regression, Lasso, and other ℓ_p penalties

Note: All of these are linear. The solutions are hyperplanes. The different methods differ only in how they *place* the hyperplane.

Ridge regression

Suppose we obtain two training samples $\mathcal{X}_1$ and $\mathcal{X}_2$ from the same distribution.

- Ideally, the linear regression solutions on both should be (nearly) identical.
- With standard linear regression, the problem may not be solvable (if $\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}$ not invertible).
- Even if it is solvable, if the matrices $\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}$ are close to singular (small spectral condition $c(\tilde{\mathbf{X}}^t\tilde{\mathbf{X}})$), then the two solutions can differ significantly.
- Ridge regression stabilizes the inversion of $\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}$. Consequences:
  - Regression solutions for $\mathcal{X}_1$ and $\mathcal{X}_2$ will be almost identical if λ sufficiently large.
  - The price we pay is a bias that grows with λ.

Lasso

- The ℓ_1-costraint "switches off" dimensions; only some of the entries of the solution $\boldsymbol{\beta}^{\text{lasso}}$ are non-zero (sparse $\boldsymbol{\beta}^{\text{lasso}}$).
- This variable selection also stabilizes $\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}$, since we are effectively inverting only along those dimensions which provide sufficient information.
- No closed-form solution; use numerical optimization.

Formulation as optimization problem

Method	$f(\boldsymbol{\beta})$	$g(\boldsymbol{\beta})$	Solution method
Least squares	$\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2$	0	Analytic solution exists if $\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}$ invertible
Ridge regression	$\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2$	$\|\boldsymbol{\beta}\|_2^2 - t$	Analytic solution exists
Lasso	$\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2$	$\|\boldsymbol{\beta}\|_1 - t$	Numerical optimization

Model Bias and Variance

Overview

- We have already encountered the fact that we can trade off model flexibility against stability of estimates (e.g. shrinkage).
- To make this effect a bit more precise, we have to discuss the type of errors that we encounter in estimation problems.
- In this context, it is useful to interpret models as sets of probability distributions.

Not examinable.

The space of probability measure

We denote the set of probability distributions on $\mathbf{X}$ by $\mathbf{M}(\mathbf{X})$.

Example: $\mathbf{X} = \{a, b, c\}$

- We write $\delta_{\{a\}}$ for the distribution with

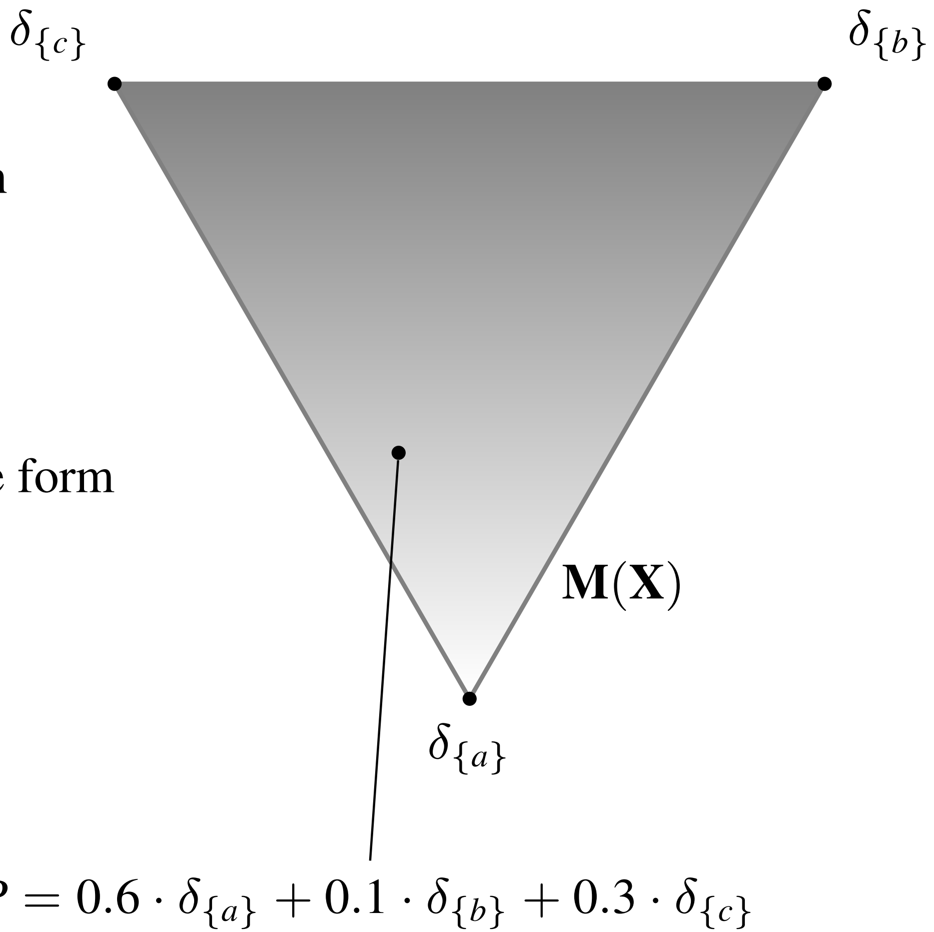
$$\Pr\{X = a\} = 1 \;,$$

  similarly for b and c.

- Every distribution $P \in \mathbf{M}(\mathbf{X})$ is of the form

$$P = c_a\delta_{\{a\}} + c_b\delta_{\{b\}} + c_c\delta_{\{c\}}$$

  with $c_1 + c_2 + c_3 = 1$.



Not examinable.

Point Masses

Dirac distributions

A **Dirac distribution** δ_x is a probability distribution which concentrates all its mass at a single point x. A Dirac δ_x is also called a **point mass**.

Note: This means that there is no uncertainty in a random variable X with distribution δ_x: We know before we even sample that $X = x$ with probability 1.

Working with a Dirac

The defining property of a Dirac is that

$$\int_{\mathbf{X}} f(x)\delta_{x_0}(dx) = f(x_0)$$

for every (integrable) function f.

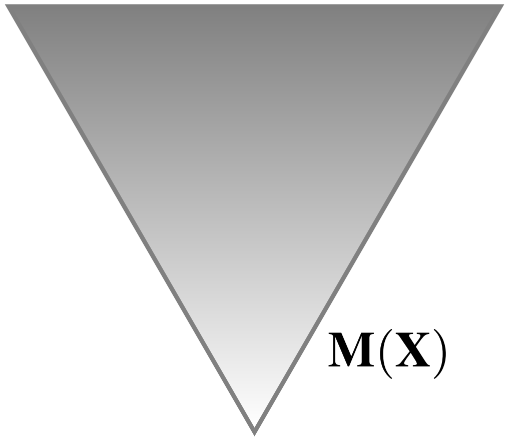
Not examinable.

$\mathbf{M}(\mathbf{X})$ for an infinite set $\mathbf{X}$

- If $\mathbf{X}$ is infinite (e.g. $\mathbf{X} = \mathbb{R}^d$), the distributions $\delta_{\{a\}}, \delta_{\{b\}}, \delta_{\{c\}}$ above are replaced by Diracs $\delta_{\mathbf{x}}$ (one for each $\mathbf{x} \in \mathbf{X}$).
- The distributions $\delta_{\mathbf{x}}$ still have the property that they cannot be represented as convex combinations.
- Hence: Each $\delta_{\mathbf{x}}$ is an extreme point of $\mathbf{M}(\mathbf{X})$.
- We need one additional dimension for each point $\mathbf{x} \in \mathbf{X}$.
- Roughly speaking, $\mathbf{M}(\mathbf{X})$ is the infinite-dimensional analogue of a triangle or tetraeder, with its extreme points labelled by the points in $\mathbf{X}$.

Visualization

In the following, we will still visualize $\mathbf{M}(\mathbf{X})$ as a triangle, but keep in mind that *this is a cartoon.*

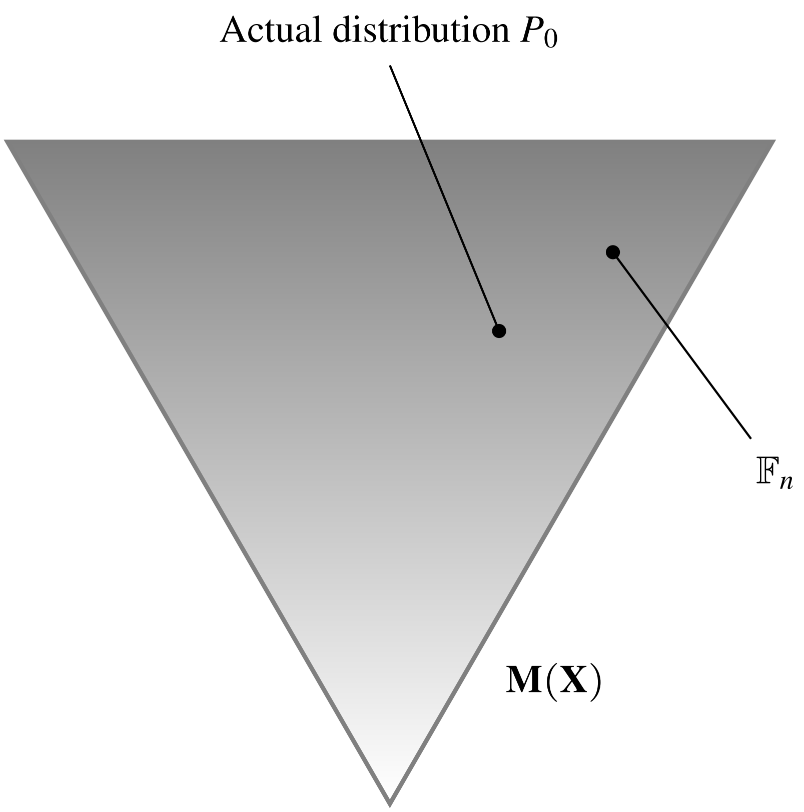


Not examinable.

The empirical distribution

If $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is a sample, its empirical distribution is

$$\mathbb{F}_n := \sum_{i=1}^{n} \frac{1}{n} \delta_{\mathbf{x}_i} \ .$$



The sample as a distribution

Using $\mathbb{F}_n$, we can regard the sample as an element of the space $\mathbf{M}(\mathbf{X})$.

For i.i.d. samples, the law of large numbers says that $\mathbb{F}_n$ converges to the true distribution as $n \to \infty$.

Not examinable.