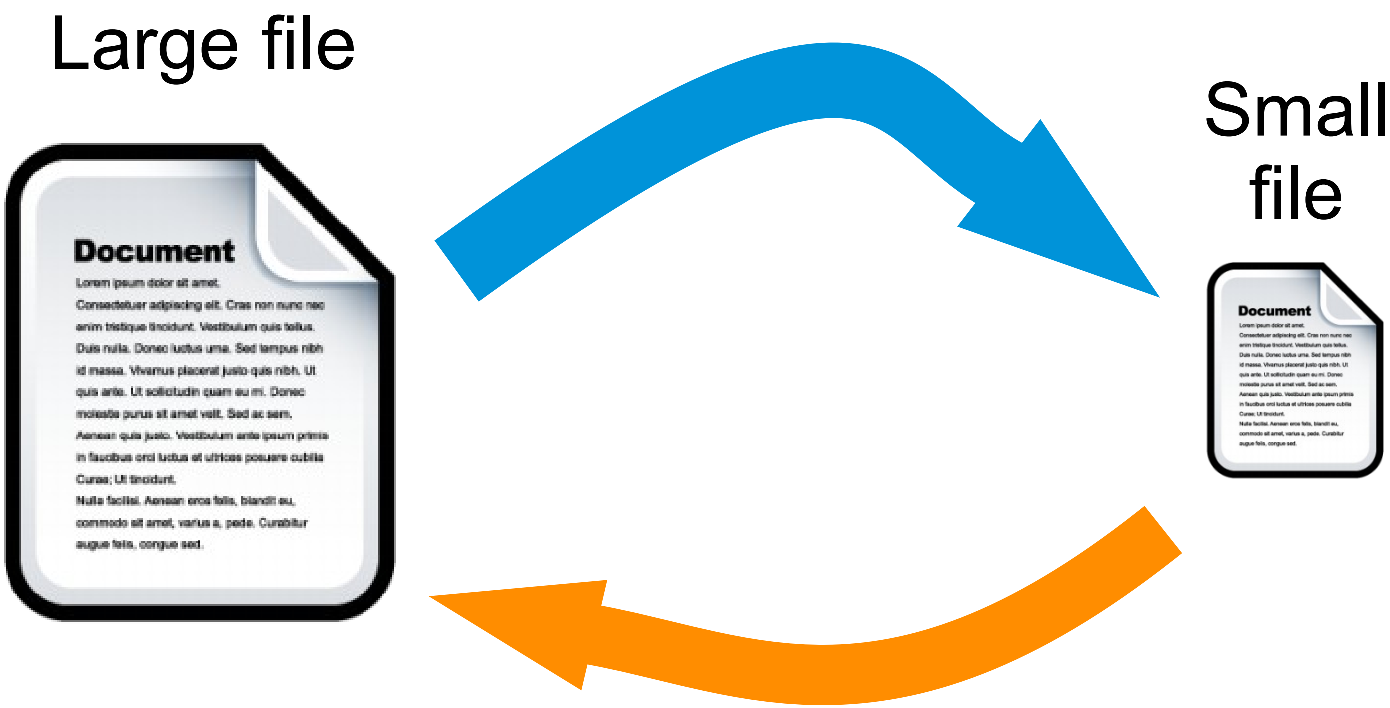
ERCIM, 2013

Minimum Description Length from a Frequentist's Perspective

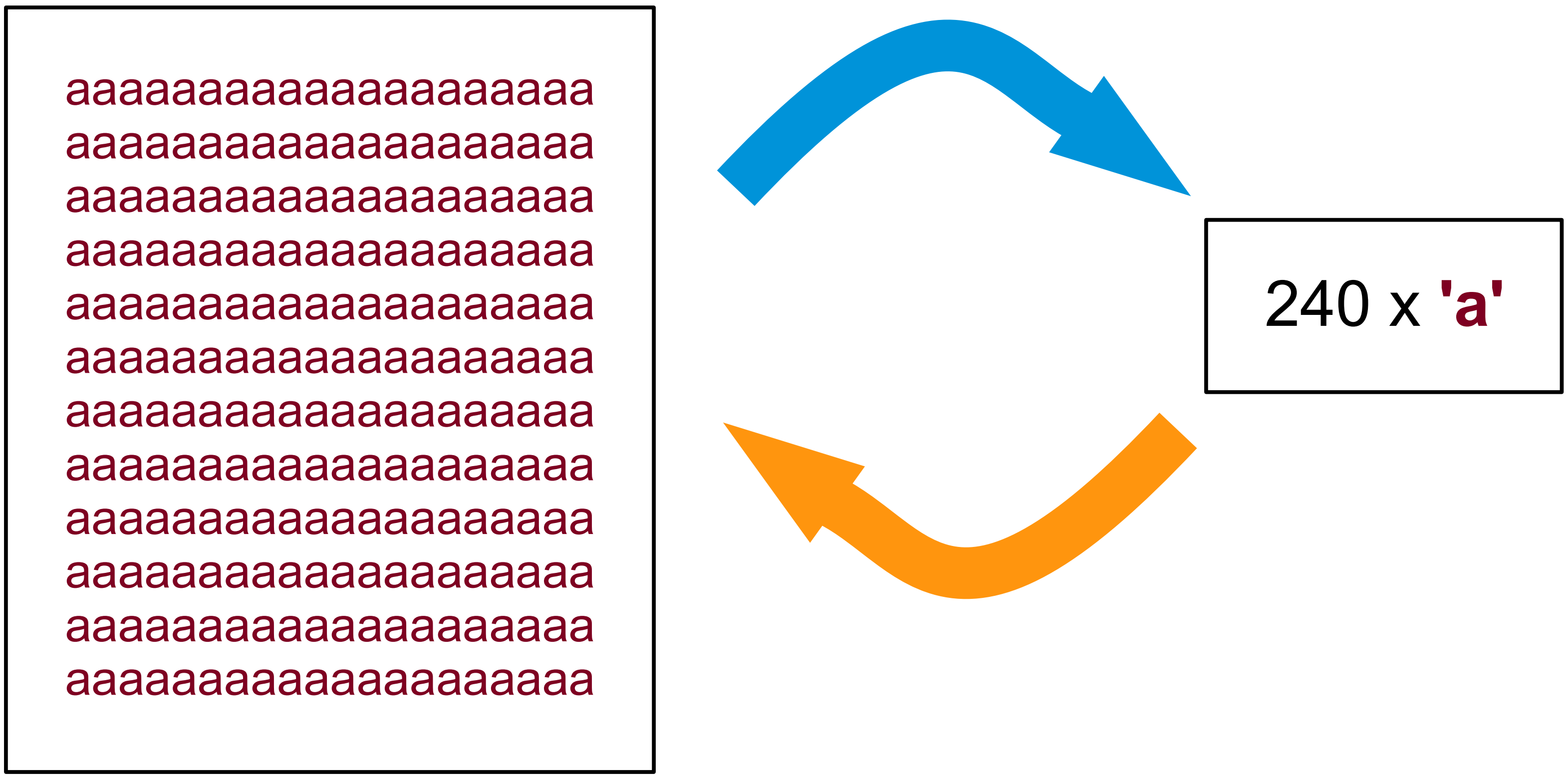
Tim van Erven

Data Compression

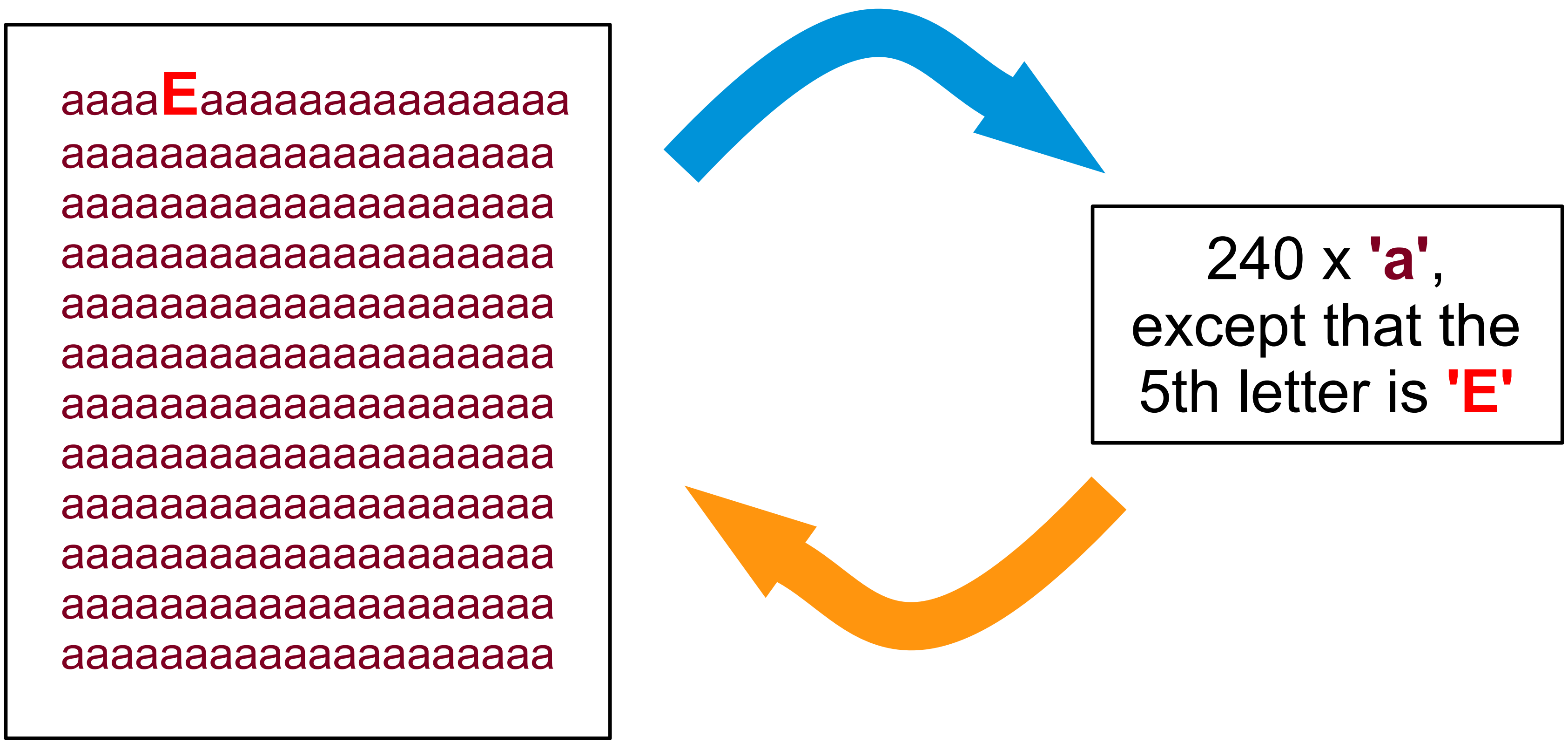


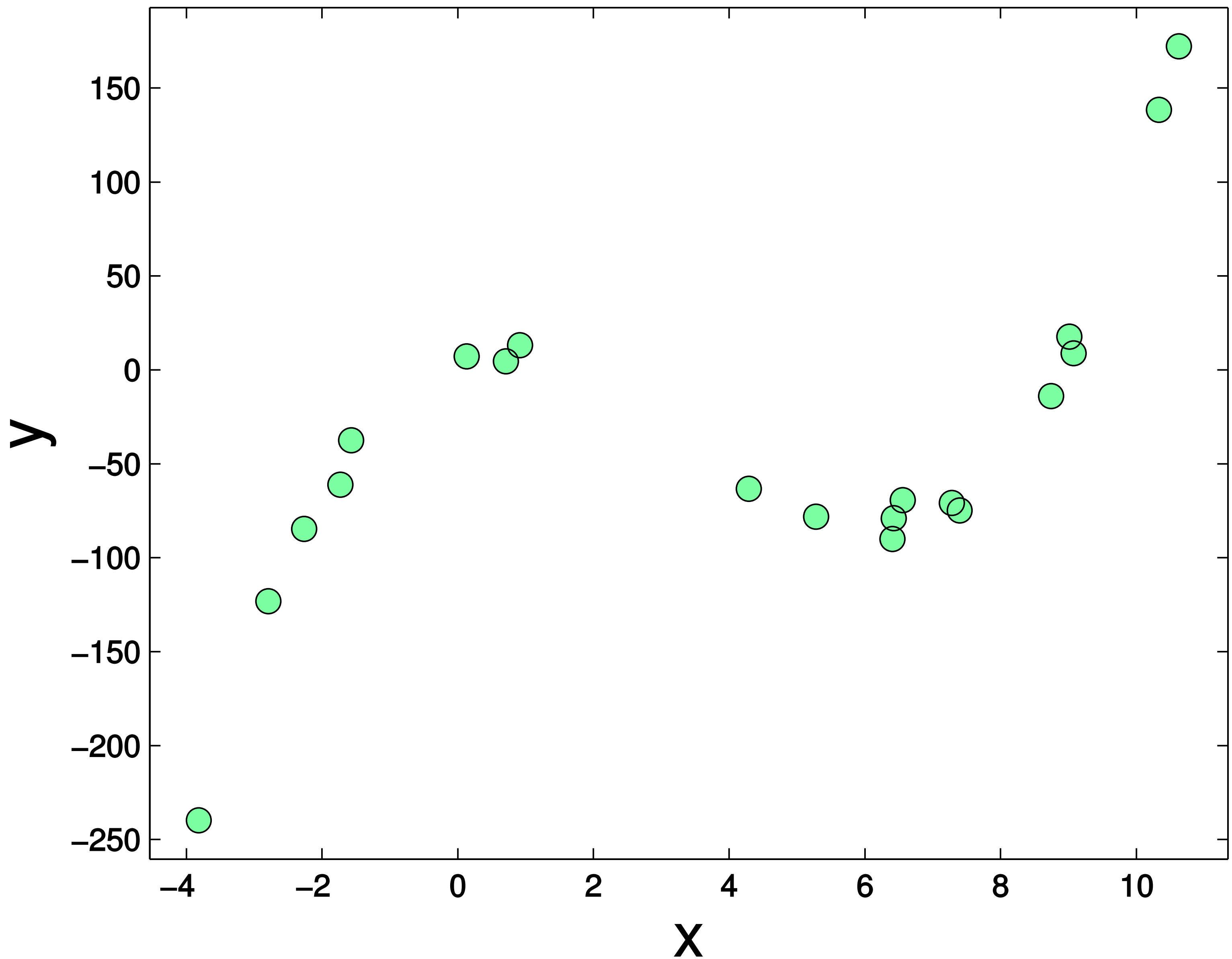
For example with WinZip

How Does Data Compression Work?

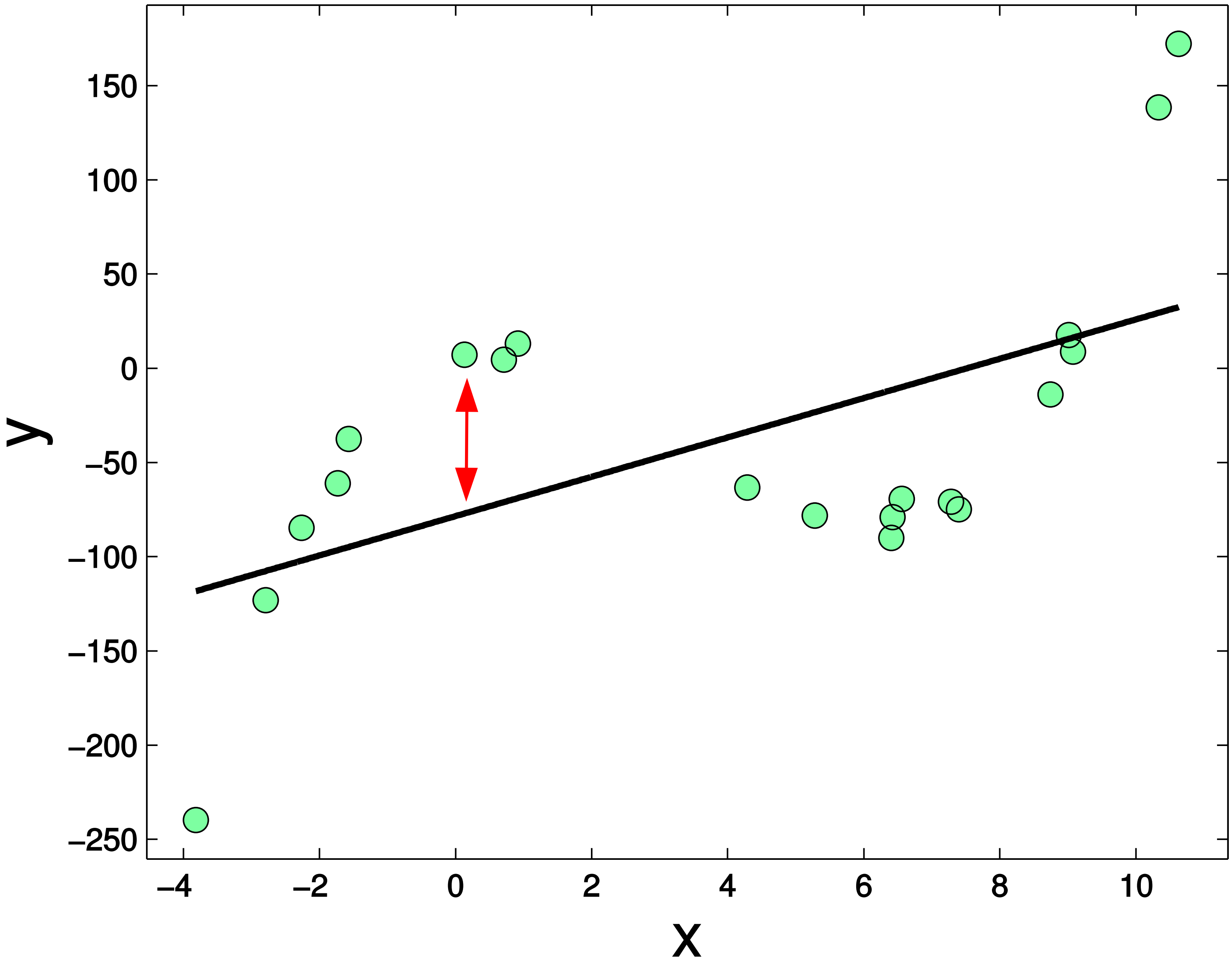


Deviations from Pattern



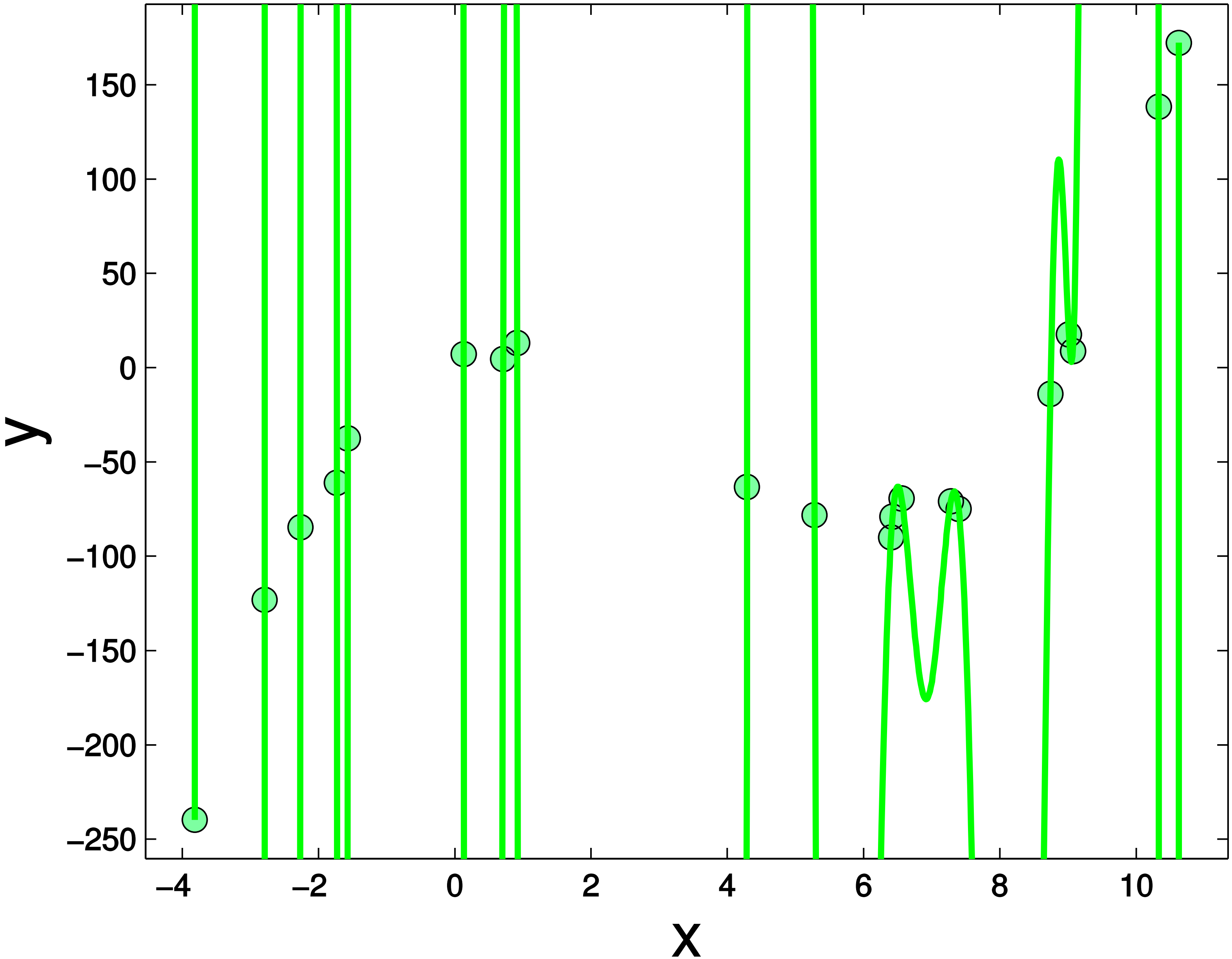


$n = 20$

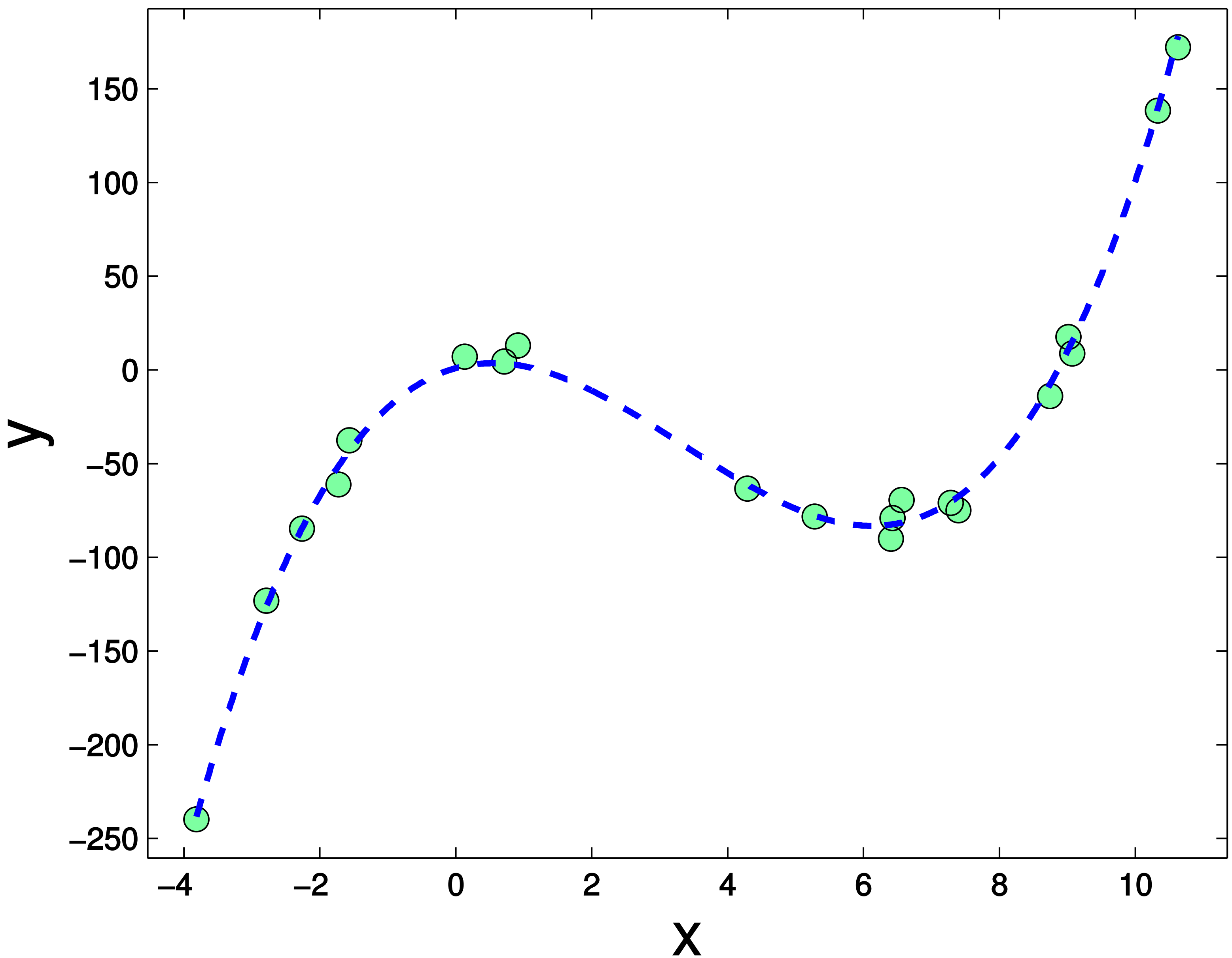


$$y = \theta_1 x + \theta_0$$

$$y = \theta_{19} x^{19} + \ldots + \theta_1 x + \theta_0$$



$$y = \theta_{19}x^{19} + \ldots + \theta_1 x + \theta_0$$



$$y = x^3 - 10x^2 + 10x + 1 + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, 5)$$

$$x \sim \mathrm{unif}[-4, 11]$$

Compressing Regression Data

- First describe **coefficients** $\theta = (\theta_0, \ldots, \theta_d)$ of polynomial

$$L(\theta) = \sum_{i=0}^{d} O(\log(n|\theta_i|))$$

- Then how the data **deviate** from the polynomial

$$L_\theta(D) = O\Big(\sum_{i=1}^{n} (y_i - f_\theta(x_i))^2\Big)$$

- If polynomial of small degree d gives small errors $\rightarrow$ good compression

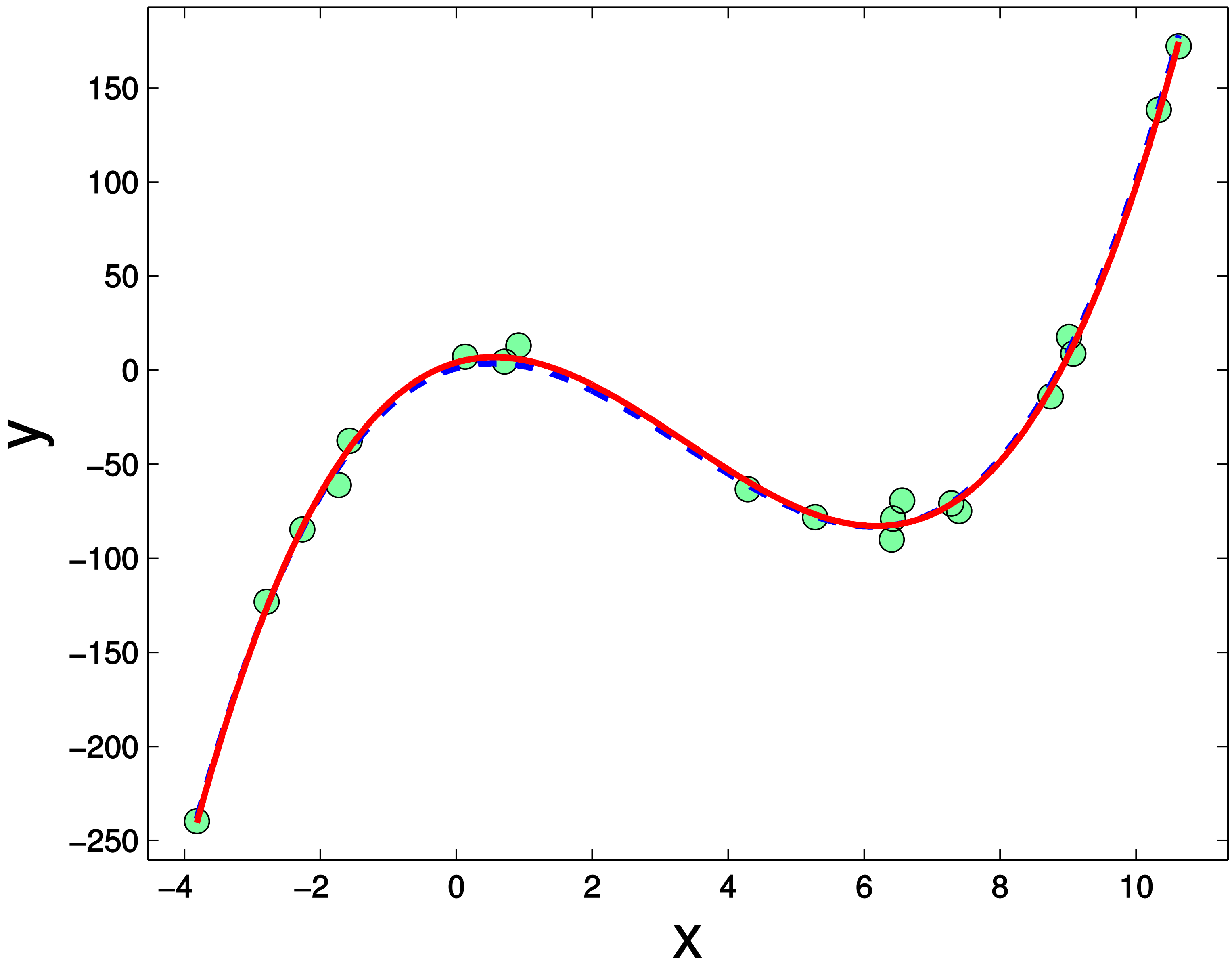
Minimum Description Length

- MDL: $$\min_{\theta} \left\{ L(\theta) + L_{\theta}(D) \right\}$$

- $L(\theta)$ = length of description of **coefficients** of polynomial, increases with degree of polynomial
- $L_{\theta}(D)$= proportional to the **errors** of the polynomial, decreases with degree of polynomial

MDL trades off degree with fit on the data!

MDL selects correct order



MDL in General

- Statistical model $\mathcal{M} = \{P_1, P_2, \ldots\}$ for data D
- **Regularity**: $L(P) = -\log \pi(P)$
  - where π is a prior distribution on $\mathcal{M}$
    (there are detailed guidelines for choosing π)
- **Deviations from pattern**: $L_P(D) = -\log P(D)$
- MDL: $$\min_{P \in \mathcal{M}} \Big\{ L(P) + L_P(D) \Big\}$$
  $$= \min_{P \in \mathcal{M}} \Big\{ -\log \pi(P) - \log P(D) \Big\}$$

Data Compression = Statistics... Almost!

- Modified MDL: $\min_P \left\{ \mathbf{2} L(P) + L_P(D) \right\}$

Thm Barron&Cover,1991: If $R_n(Q) \to 0$, then the **modified MDL** estimator converges:

$$\mathrm{Hel}^2(Q, \hat{P}) \lesssim R_n(Q) \qquad \text{in probability}$$

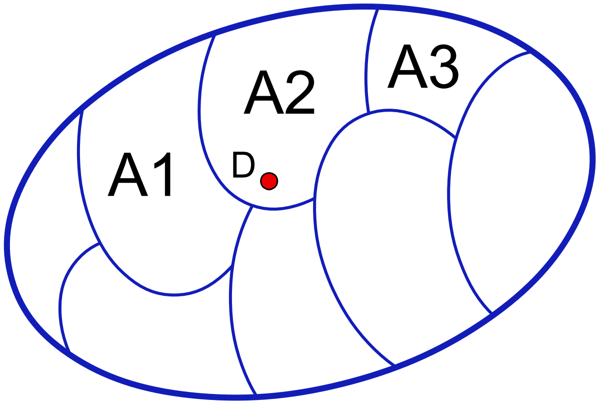
as $n \to \infty$.

- IID data: $D = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} Q$
- Rate is minimum **expected description length**:

$$R_n(Q) = \min_P \left\{ \frac{2L(P)}{n} + \mathrm{KL}(Q \| P) \right\}$$

Standard MDL Can Go Wrong

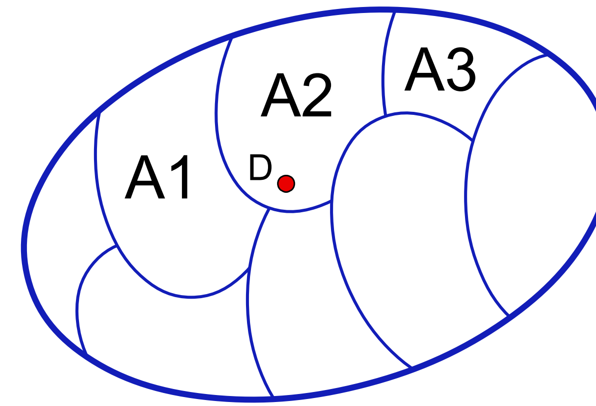
- $\mathcal{M} = \{Q, P_1, P_2, \ldots\} \quad \pi(Q) = \frac{1}{3}$
- $P_i(D) = Q(D|A_i) \qquad \pi(P_i) = \frac{2}{3} Q(A_i)$



Partition of sample space

Standard MDL Can Go Wrong

- $\mathcal{M} = \{Q, P_1, P_2, \ldots\} \quad \pi(Q) = \frac{1}{3}$
- $P_i(D) = Q(D|A_i) \qquad \pi(P_i) = \frac{2}{3}Q(A_i)$



Partition of sample space

- Then $\sum_i \pi(P_i)P_i(D)$ "looks like" Q and MDL gets confused:

$$L(Q) + L_Q(D) = \log 3 - \log Q(D)$$

$$L(P_i) + L_{P_i}(D) = \log \frac{3}{2} - \log Q(D) \qquad \text{if } D \in A_i.$$

- (In this example Bayes posterior does not converge either, so Bayesian parameter estimation is in trouble too.)

But Bad Example Can Be Excluded

- Problem in a nutshell: if $D \in A_i$, then

$$\pi(P_i)P_i(D) \approx \pi(Q)Q(D)$$
$$-\log Q(D) + \log P_i(D) \approx L(P_i) - L(Q)$$

- A_i = set where P_i has all its mass

But Bad Example Can Be Excluded

- Problem in a nutshell: if $D \in A_i$, then

$$\pi(P_i)P_i(D) \approx \pi(Q)Q(D)$$
$$-\log Q(D) + \log P_i(D) \approx L(P_i) - L(Q)$$

- A_i = set where P_i has all its mass
- Expected version of **problematic distributions**:

$$\mathcal{A}_n = \left\{ P \in \mathcal{M} \mid n\,\mathrm{KL}(P\|Q) \approx L(P) - L(Q) \right\}$$

But Bad Example Can Be Excluded

- Problem in a nutshell: if $D \in A_i$, then

$$\pi(P_i)P_i(D) \approx \pi(Q)Q(D)$$
$$-\log Q(D) + \log P_i(D) \approx L(P_i) - L(Q)$$

- A_i = set where P_i has all its mass
- Expected version of **problematic distributions**:

$$\mathcal{A}_n = \left\{ P \in \mathcal{M} \mid n \, \mathrm{KL}(P \| Q) \approx L(P) - L(Q) \right\}$$

$$\mathcal{A}_n = \left\{ P \mid nc_1 D_\alpha(P \| Q) < L(P) - L(Q) < nc_2 D_\beta(P \| Q) \right\}$$

where D_α is Rényi divergence, $\alpha < 1 < \beta$, $D_1 = \mathrm{KL}$

Negligible Set Condition

- **Negligible set condition**: the set of problematic densities

$$\mathcal{A}_n = \Big\{ P \mid n\,\mathrm{KL}(P \| Q) \approx L(P) - L(Q) \Big\}$$

has **small prior probability**:

$$\pi \Big\{ P \in A_n \mid \mathrm{Hel}^2(P, Q) \geq \epsilon \Big\} \leq a e^{-bn\epsilon} \qquad \text{for all } \epsilon > 0.$$

Standard MDL Does Work

Thm Van Erven, 2010: If the **negligible set condition** holds and $R_n(Q) \to 0$, then the **standard MDL** estimator converges:

$$\mathrm{Hel}^2(Q, P_{\hat{\theta}}) \lesssim R_n(Q) \qquad \text{in probability}$$

as $n \to \infty$.

- Rate is minimum **expected description length**:

$$R_n(Q) = \min_P \left\{ \frac{L(P)}{n} + \mathrm{KL}(Q \| P) \right\}$$

Understanding Modified MDL

Modified MDL: $\min_{P} \left\{ \mathbf{2} L(P) + L_P(D) \right\}$

Lemma (Van Erven, 2010):
For **modified MDL**, the negligible set condition is **automatically satisfied**.

Summary

- Can use data compression (MDL) to fit parameters and prevent overfitting
- **Works well if modified** with weird factor of 2, which makes no sense for data compression
- New results:
  - Works if **problematic distributions** A_n have small prior probability (otherwise counterexample)
  - Factor 2 is a simple way to guarantee this.
- **Understanding of MDL from a frequentist perspective**

Future Work

- Do we need to add the factor of 2 **in practice**?
  - Problematic distributions seem pretty pathological
  - Practitioners use MDL without factor of 2 without problems

References

- Barron, Cover. Minimum complexity density estimation. IEEE Transactions on Information Theory, 37(4):1034-1054, 1991.
- Van Erven, When Data Compression and Statistics Disagree. **PhD thesis**, Leiden University, 2010. **Chapter 5**.
- Van Erven, Harremoës, Rényi Divergence and Kullback-Leibler Divergence. To appear in IEEE Transactions on Information Theory, 2014.

Back to Modified MDL

- If there exists $L'(P)$ s.t. $L(P) = 2L'(P) + C$ then standard MDL with $L(P)$ = modified MDL with $L'(P)$.

Lemma (B&C, 1991): there exists $L'(P)$ such that $L(P) = 2L'(P) + C$ if and only if the **light tails condition**

$$\sum_P \pi(P)^{1/2} \leq B < \infty$$

holds.

Proof: Take $L'(P) = -\log \frac{\pi(P)^{1/2}}{B}, C = -2 \log B$

Lemma (Van Erven, 2010): Light tails **implies negligible set condition!**