Toric Modification on Machine Learning

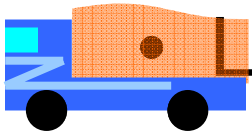
Keisuke Yamazaki & Sumio Watanabe

Tokyo Institute of Technology

Agenda

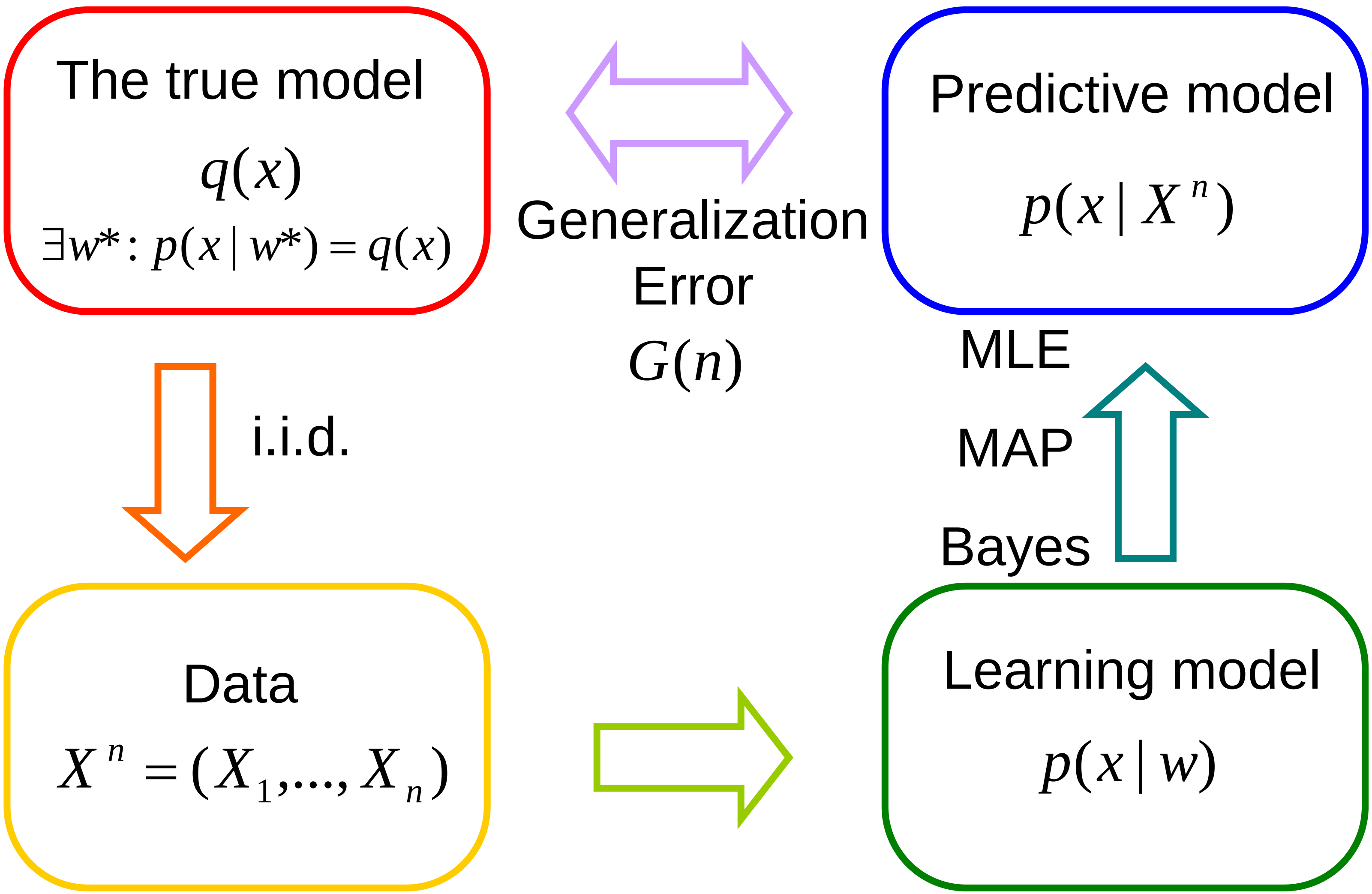
- Learning theory and algebraic geometry
- Two forms of the Kullback divergence
- Toric modification
- Application to a binomial mixture model
- Summary

Agenda



- Learning theory and algebraic geometry
- Two forms of the Kullback divergence
- Toric modification
- Application to a binomial mixture model
- Summary

What is the generalization error?



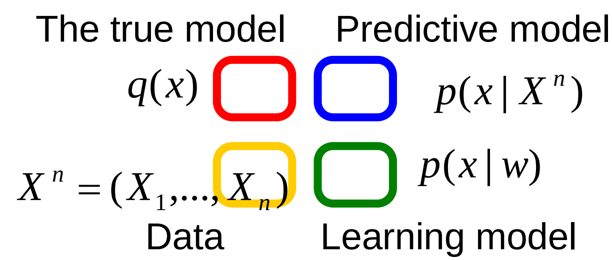
Algebraic geometry connected to learning theory in the Bayes method.

- The formal definition of the generalization error.

$$G(n) = E_{X^n}\left[\int q(x)\log\frac{q(x)}{p(x\,|\,X^n)}dx\right]$$

Another Kullback divergence :

$$H(w) = \int q(x)\log\frac{q(x)}{p(x\,|\,w)}dx$$



Algebraic Geometry

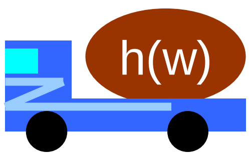
Machine Learning / Learning Theory

$G(n)$

$H(w)$

The zeta function has an important role for the connection.

Zeta function :



$$\varsigma(z) = \int h(w)^z \varphi(w) dw$$

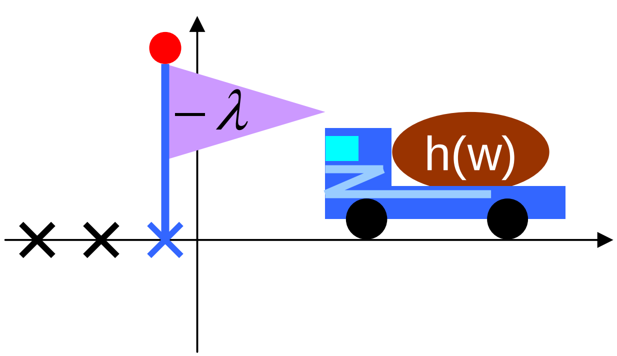
$$= \frac{f(z)}{(z+\lambda)^m} + \dots$$

$h(w)$: Analytic func.

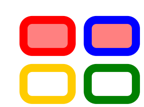
$\varphi(w)$: C-infinity func. with compact support

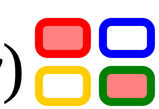
$f(z)$: holomorphic function

Algebraic Geometry



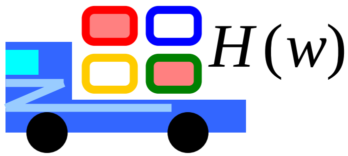
Machine Learning / Learning Theory

$G(n)$ 

$H(w)$ 

The zeta function has an important role for the connection.

Zeta function :



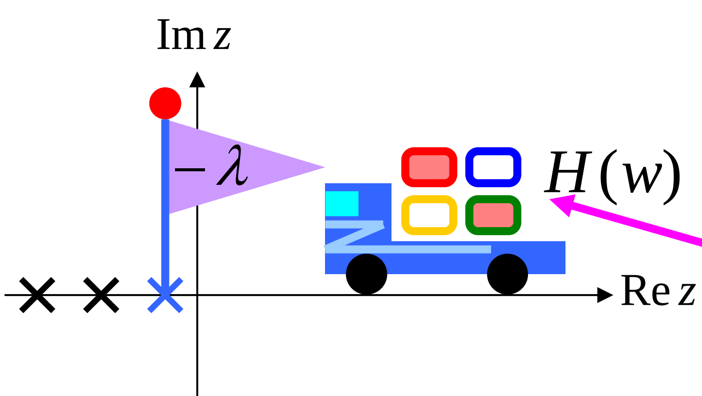
$$\varsigma(z) = \int H(w)^z \varphi(w) dw = \frac{f(z)}{(z+\lambda)^m} + \ldots$$

$H(w)$: Kullback divergence

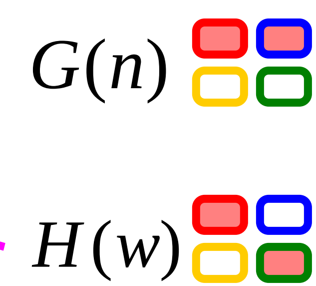
$\varphi(w)$: Prior distribution

$f(z)$: holomorphic function

Algebraic Geometry



Machine Learning / Learning Theory

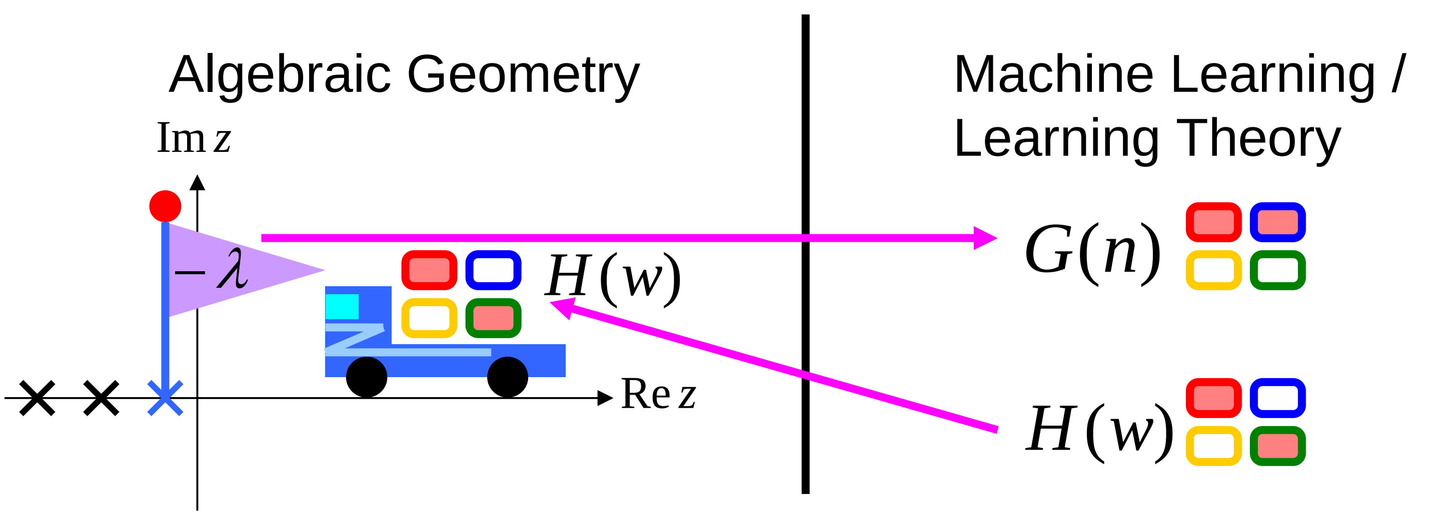


The largest pole of the zeta function determines the generalization error.

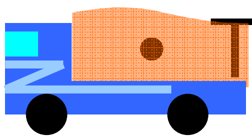
Asymptotic Bayes generalization error [Watanabe 2001]

$$G(n) = \frac{\lambda}{n} - \frac{m-1}{n \log n} + o(1/n \log n)$$

$$\varsigma(z) = \int H(w)^z \varphi(w) dw = \frac{f(z)}{(z+\lambda)^m} + \ldots$$



Agenda

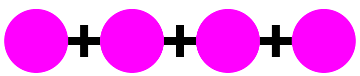
- Learning theory and algebraic geometry

- Two forms of the Kullback divergence
- Toric modification
- Application to a binomial mixture model
- Summary

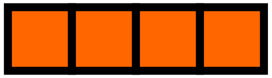
Calculation of the zeta function requires well-formed H(w).

$$H(w) = \int q(x) \log \frac{q(x)}{p(x\,|\,w)}\,dx$$

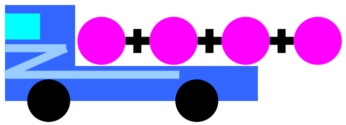
$$= (w_1 w_2 + w_3 w_4)^2 + (w_1 w_2^3 + w_3 w_4^3)^2$$

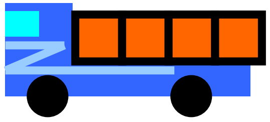
$$H(g(u)) = f(u)\,u_1^{2\alpha_1} u_2^{2\alpha_2} u_3^{2\alpha_3} u_4^{2\alpha_4}$$

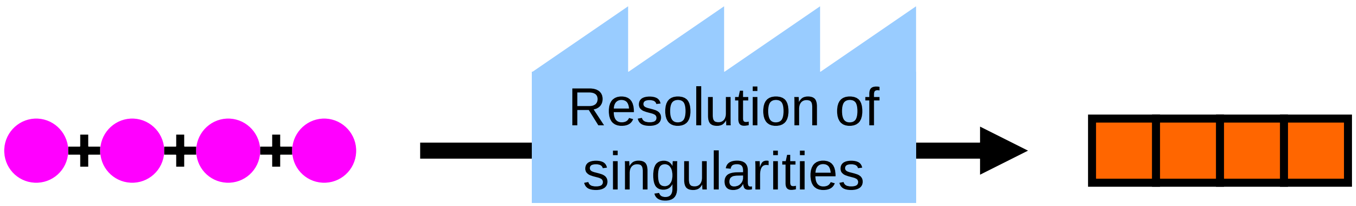

: Polynomial form


: Factorized form

$$\varsigma(z) = \int H(w)^z\,dw = \int H(g(u))^z\,|\,g'(u)\,|\,du$$



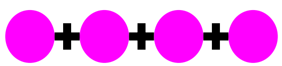




$$w = g(u)$$

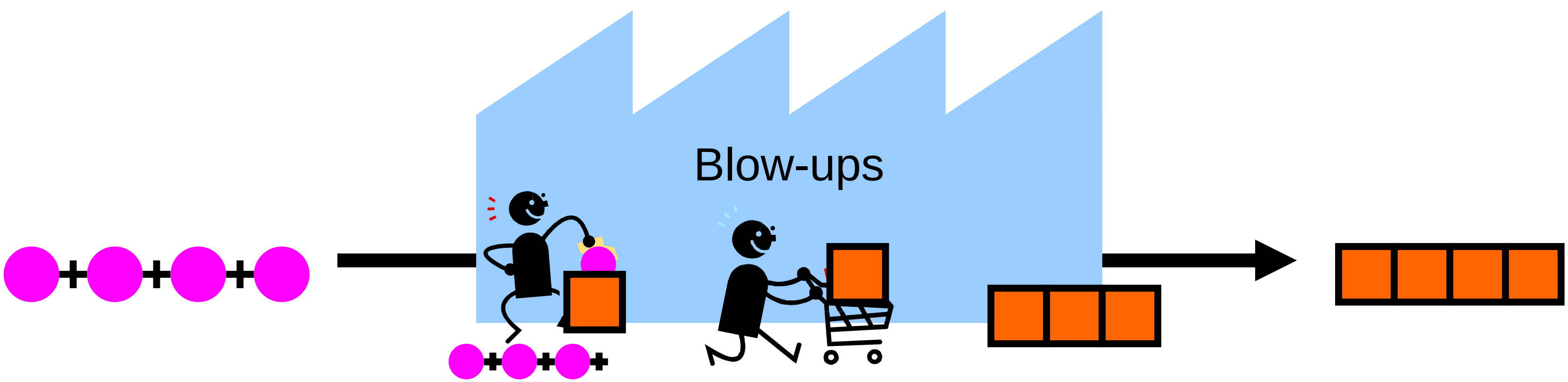
The resolution of singularities with blow-ups is an iterative method.

$$H(w)=(w_1w_2+w_3w_4)^2$$



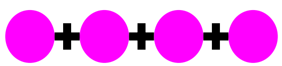
$$\begin{cases} w_1 = u_3u_1 \\ w_3 = u_3 \end{cases} \rightarrow H(w)=u_3^2(u_1w_2+w_4)^2$$

$$\begin{cases} w_1 = u_1 \\ w_3 = u_1u_3 \end{cases} \rightarrow H(w)=u_1^2(w_2+u_3w_4)^2$$



The resolution of singularities with blow-ups is an iterative method.

$$H(w) = (w_1 w_2 + w_3 w_4)^2$$

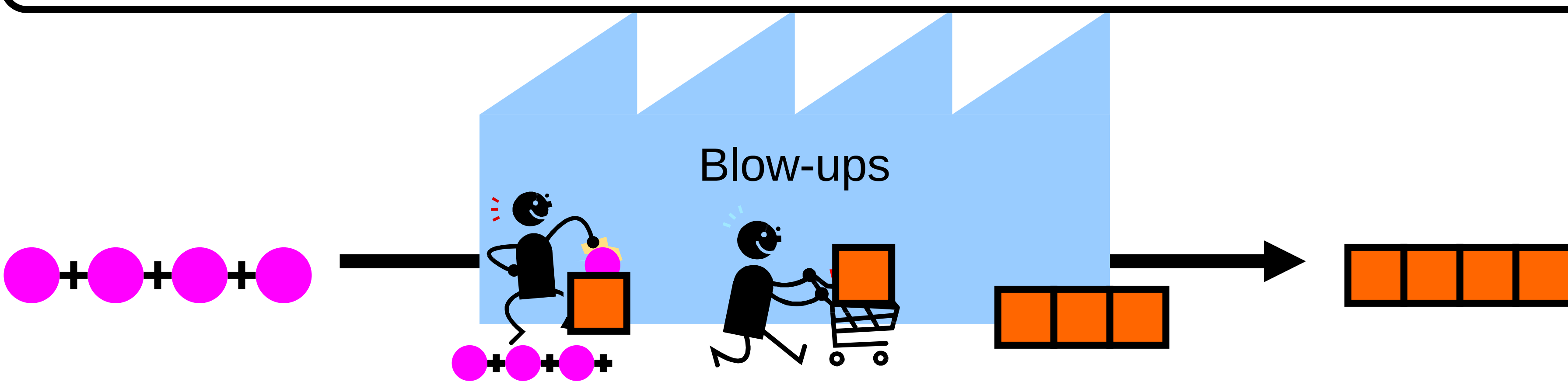


$$\begin{cases} w_1 = u_3 u_1 \\ w_3 = u_3 \end{cases} \quad \rightarrow \quad H(w) = u_3^2 (u_1 w_2 + w_4)^2$$

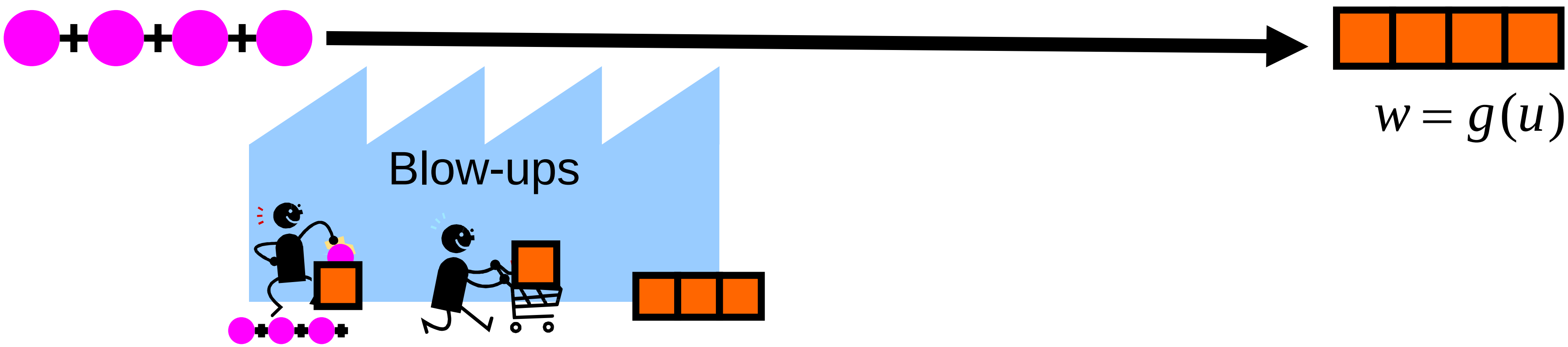
$$\begin{cases} w_1 = u_1 \\ w_3 = u_1 u_3 \end{cases} \quad \rightarrow \quad H(w) = u_1^2 (w_2 + u_3 w_4)^2$$

Kullback divergence in ML is complicated and has high dimensional w :

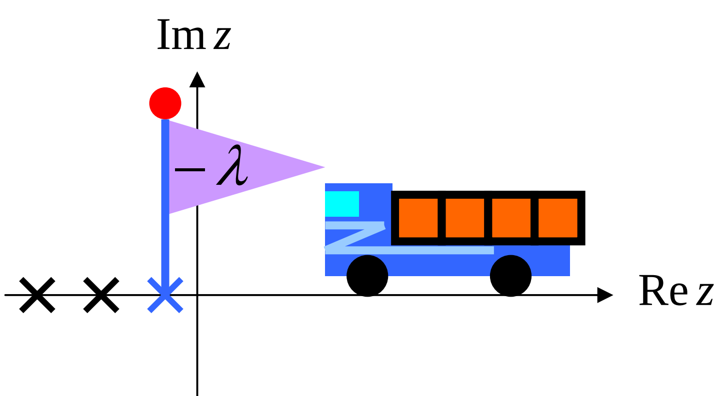
$$H(w) = (w_1 w_2 + w_3 w_4)^2 + (w_1 w_2^3 + w_3 w_4^3)^2$$



The bottleneck is the iterative method.

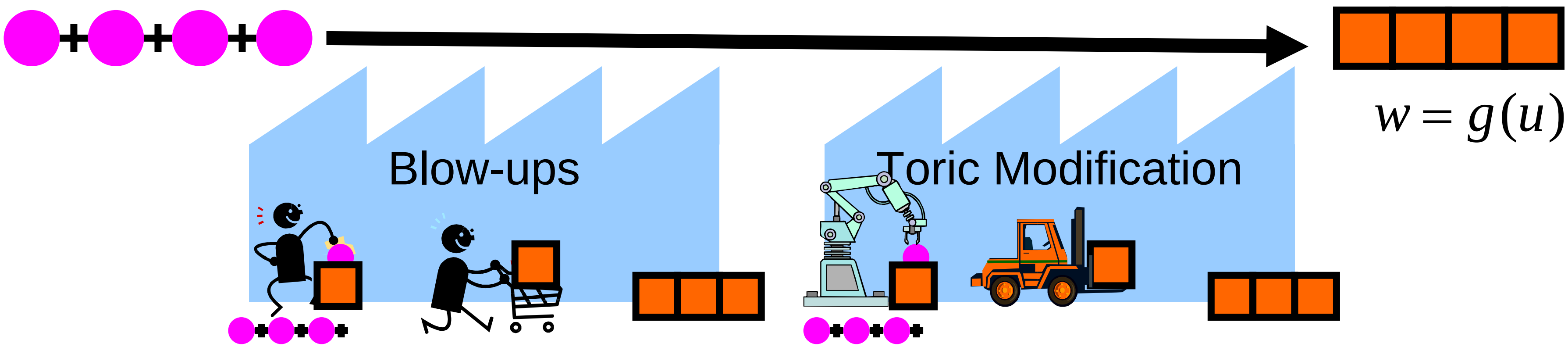


Applicable function	Any function
Computational cost	Large

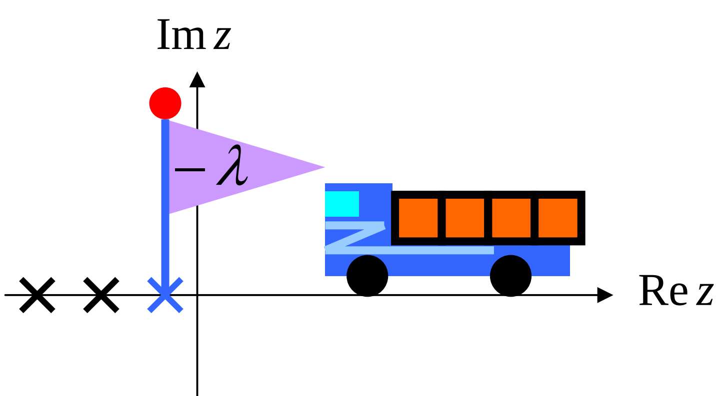


$$G(n) = \frac{\lambda}{n} - \frac{m-1}{n \log n} + o(1/n \log n)$$

Toric modification is a systematic method for the resolution of singularities.



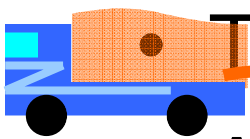
Applicable function	Any function
Computational cost	Large



$$G(n) = \frac{\lambda}{n} - \frac{m-1}{n \log n} + o(1/n \log n)$$

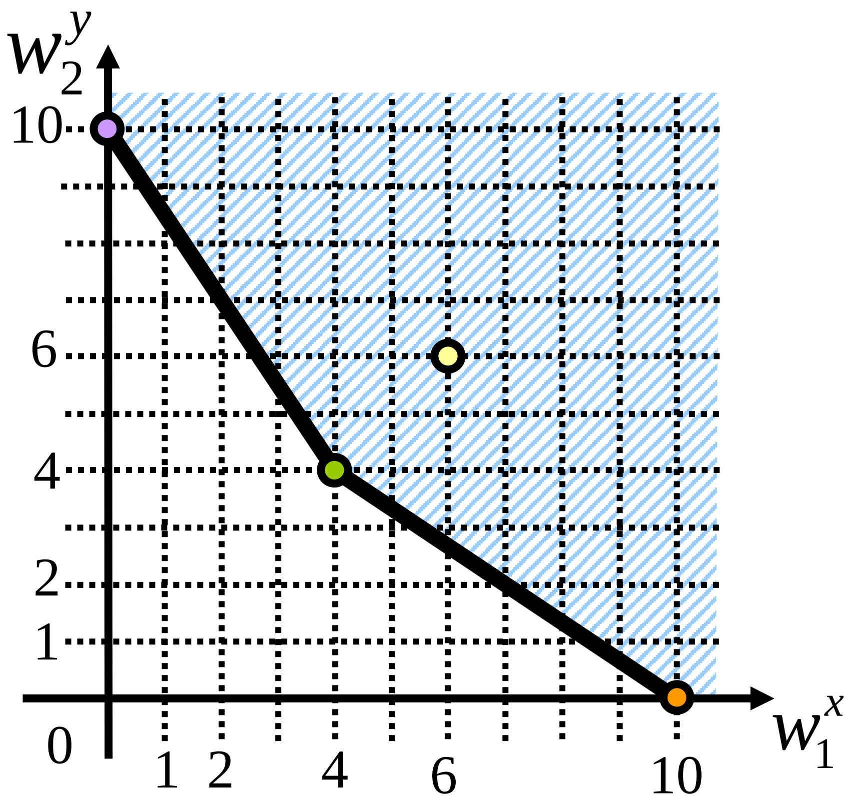
Agenda

- Learning theory and algebraic geometry
- Two forms of the Kullback divergence

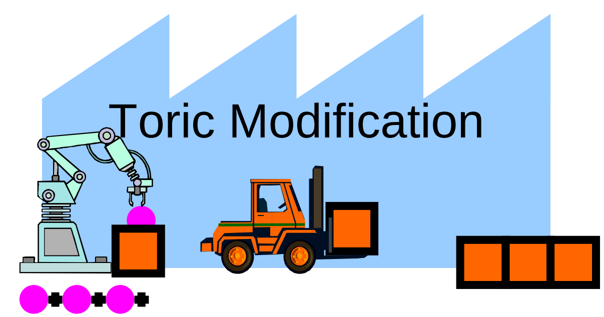


- Toric modification
- Application to a binomial mixture model
- Summary

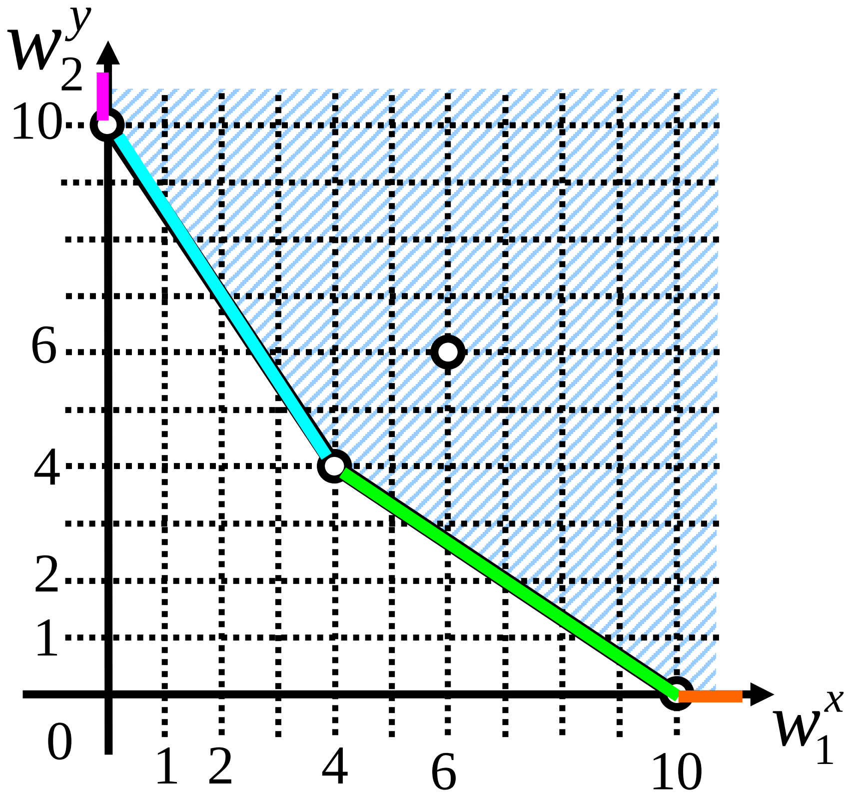
Newton diagram is a convex hull in the exponent space

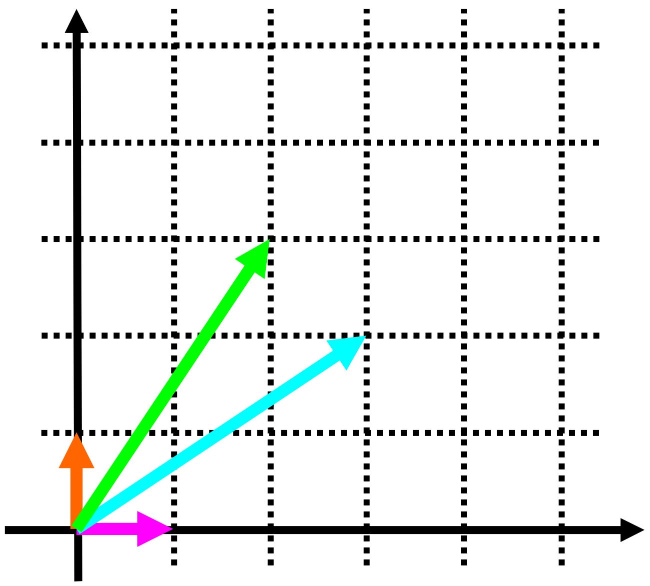


$$H(w) = w_1^{10} + w_1^{6} w_2^{6} + w_1^{4} w_2^{4} + w_2^{10}$$

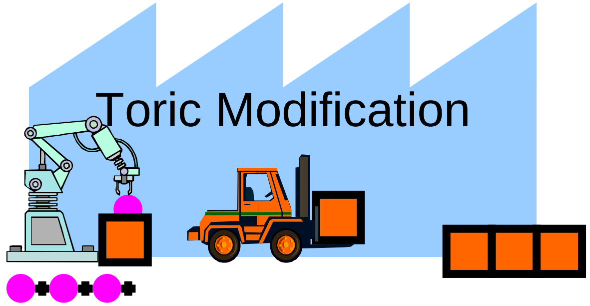


The borders determine a set of vectors in the dual space.

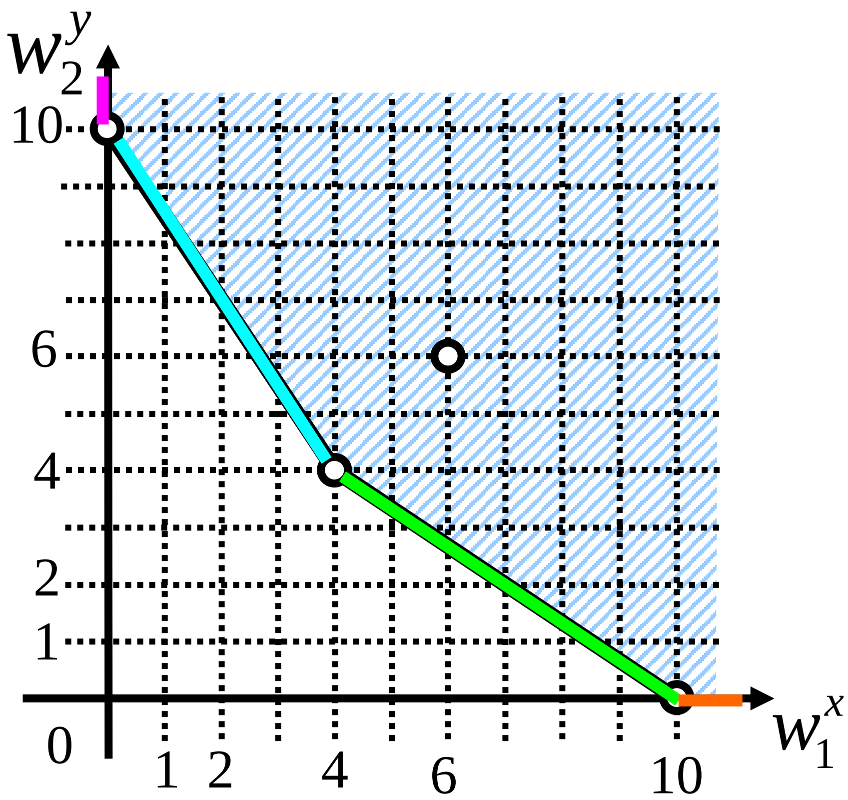


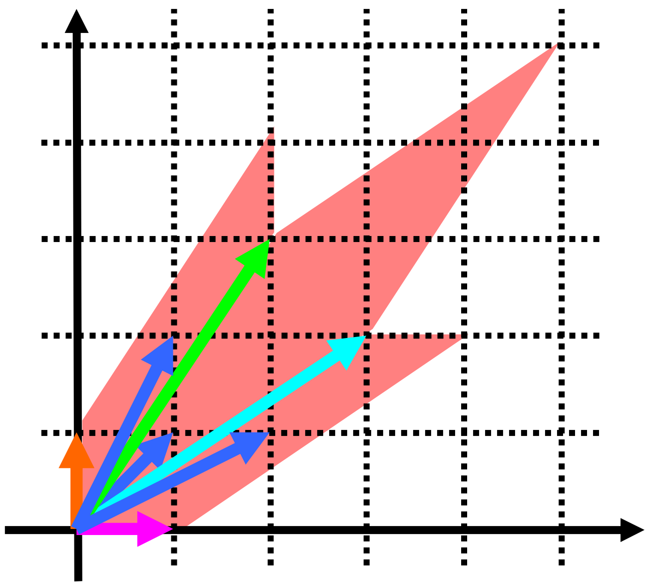


$$H(w) = w_1^{10} + w_1^6 w_2^6 + w_1^4 w_2^4 + w_2^{10}$$

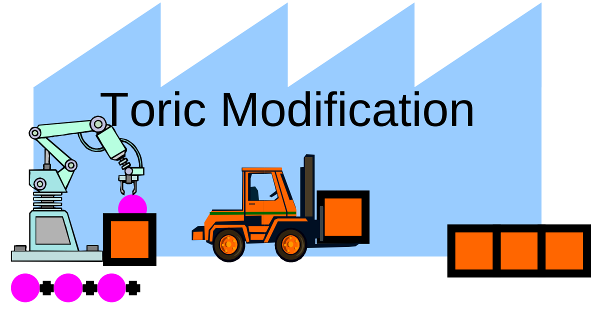


Add some vectors subdividing the spanned area.





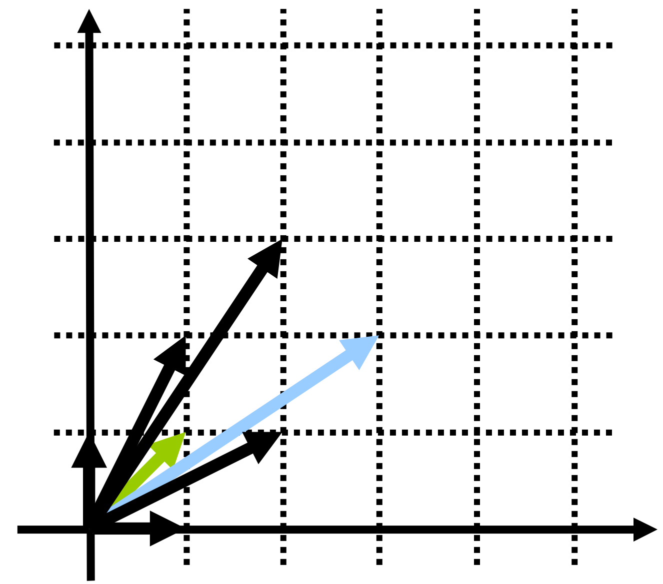
$$H(w) = w_1^{10} + w_1^6 w_2^6 + w_1^4 w_2^4 + w_2^{10}$$



Selected vectors construct the resolution map.

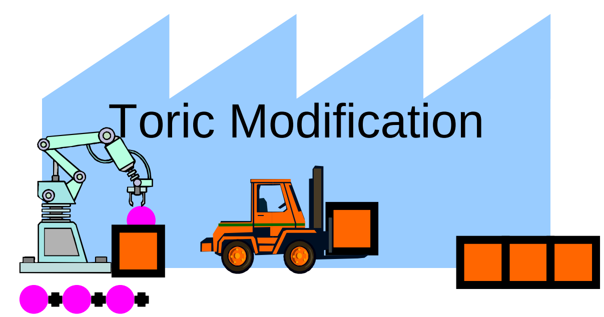
$$H(w) = w_1^{10} + w_1^6 w_2^6 + w_1^4 w_2^4 + w_2^{10}$$

$$A = \begin{pmatrix} 3 & 1 \\ 2 & 1 \end{pmatrix} \qquad s.t.\ \det A = \pm 1$$



$$w = g(u) : \begin{cases} w_1 = u_1^3 u_2^1 \\ w_2 = u_1^2 u_2^1 \end{cases}$$

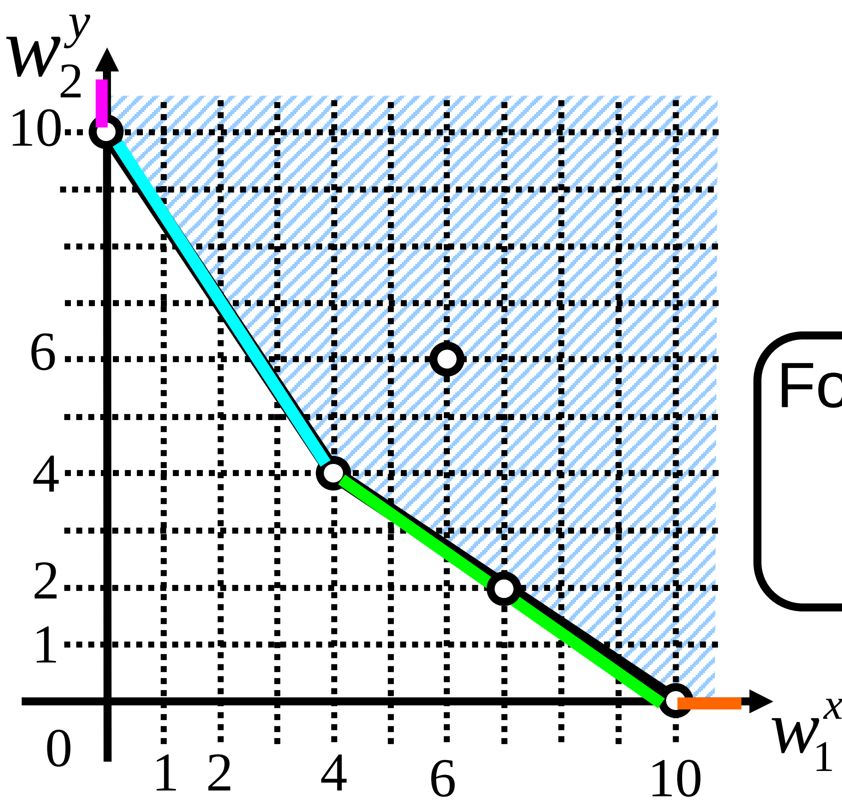
$$H(g(u)) = (u_1^{10} u_2^2 + u_1^{10} u_2^4 + 1 + u_2^2) u_1^{20} u_2^8$$



Non-degenerate Kullback divergence

- The condition to apply the toric modification
  to the Kullback divergence

$$H(w) = w_1^{10} + 2w_1^7 w_2^2 + w_1^6 w_2^6 + w_1^4 w_2^4 + w_2^{10}$$

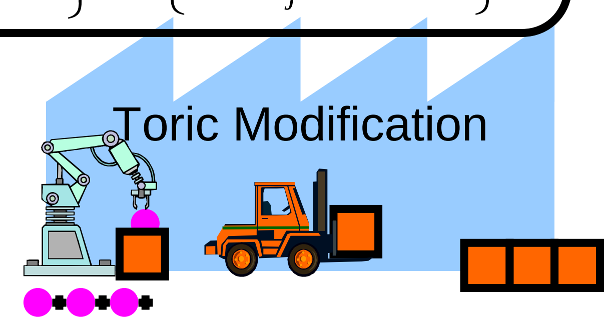


$$f_1(w) = w_1^4 w_2^4 + w_2^{10}$$

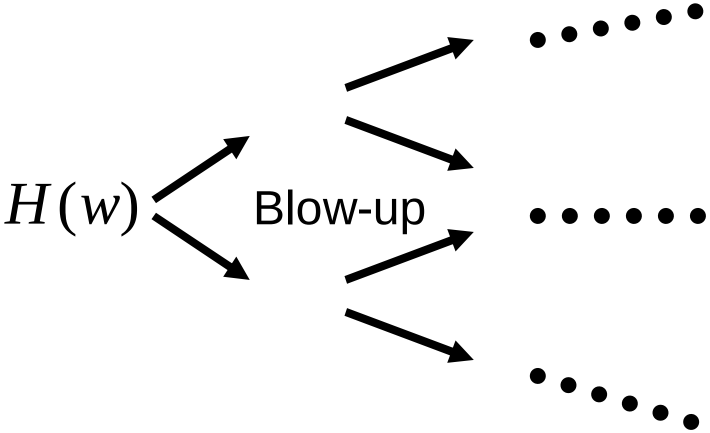
$$f_2(w) = w_1^{10} + 2w_1^7 w_2^2 + w_1^4 w_2^4$$
$$= (w_1^5 + w_1^2 w_2^2)^2$$

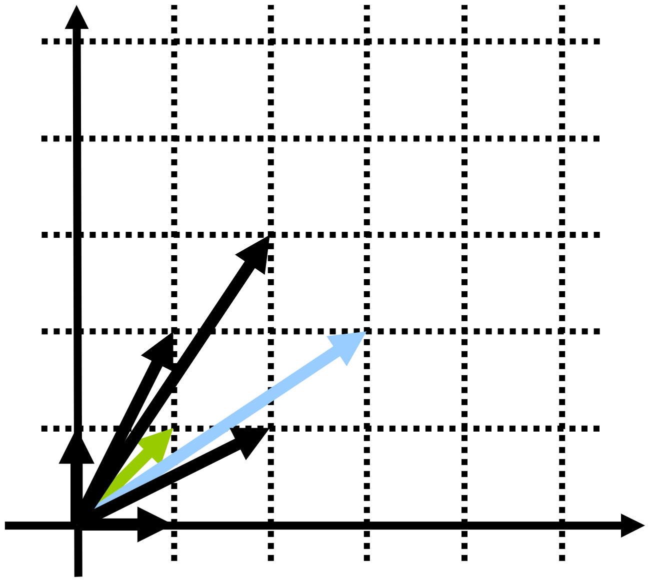
For all i :

$$\left\{ w : \frac{\partial f_i}{\partial w_j} = 0 \right\} \subset \left\{ w : \prod_j w_j = 0 \right\}$$



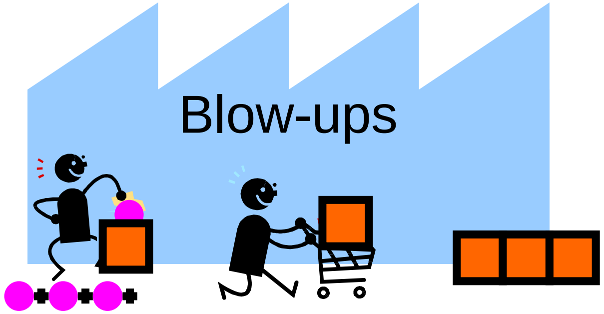
Toric modification is “systematic”.

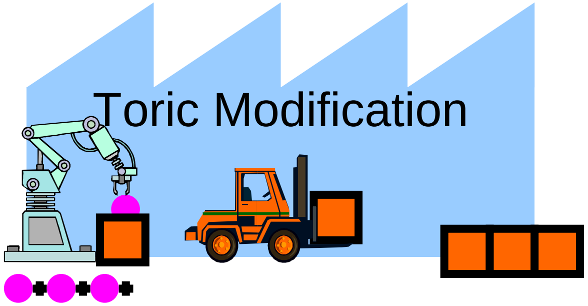




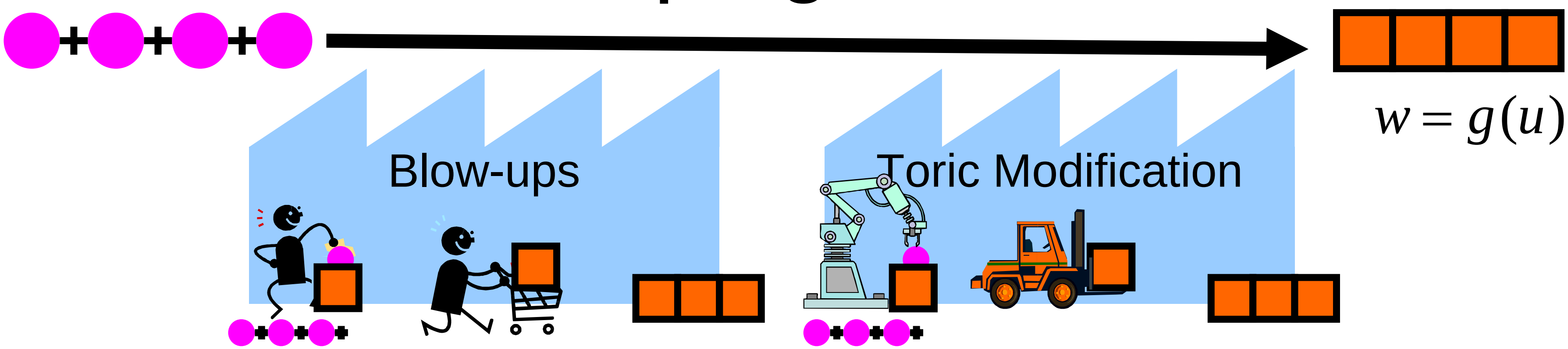
The search space will be large.
We cannot know how many iterations we need.

The number of vectors is limited.

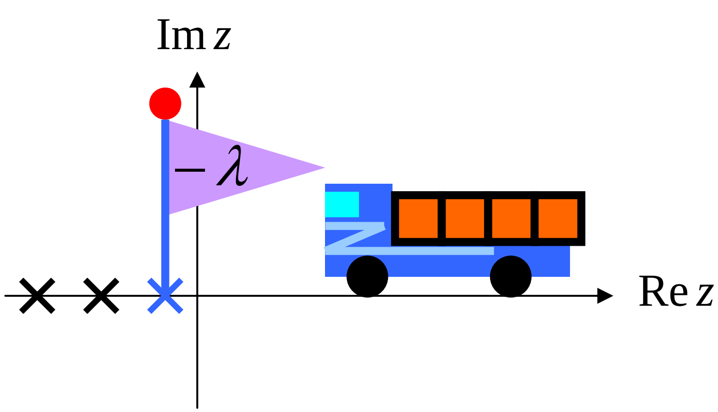




Toric modification can be an effective plug-in method.

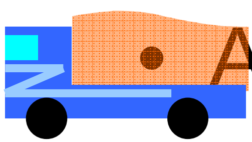


Applicable function	Any function	Non-degenerate function
Computational cost	Large	Limited



$$G(n) = \frac{\lambda}{n} - \frac{m-1}{n \log n} + o(1/n \log n)$$

Agenda

- Learning theory and algebraic geometry
- Two forms of the Kullback divergence
- Toric modification

- Application to a binomial mixture model
- Summary

An application to a mixture model

- Mixture of binomial distributions

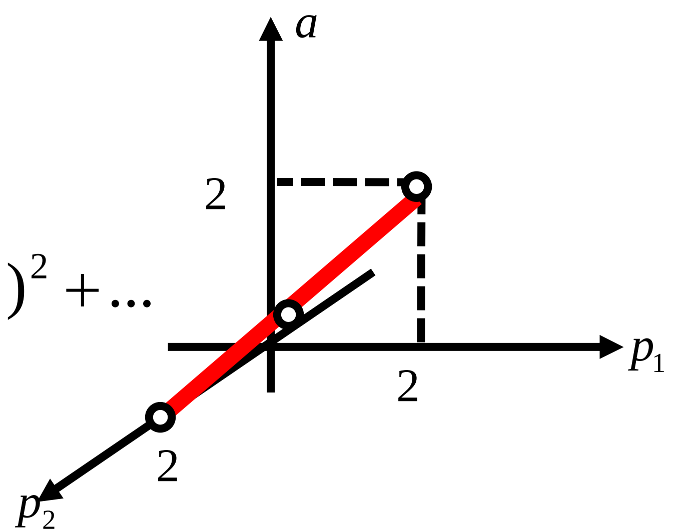
The true model: $q(x) = Bin_N(x, p^*) = \binom{N}{x} p^{*x} (1-p^*)^{N-x}$

Learning model: $p(x|w) = aBin_N(x, p_1) + (1-a)Bin_N(x, p_2)$

$$H(w) = \sum_{x=0}^{N} q(x) \log \frac{q(x)}{p(x|w)}$$

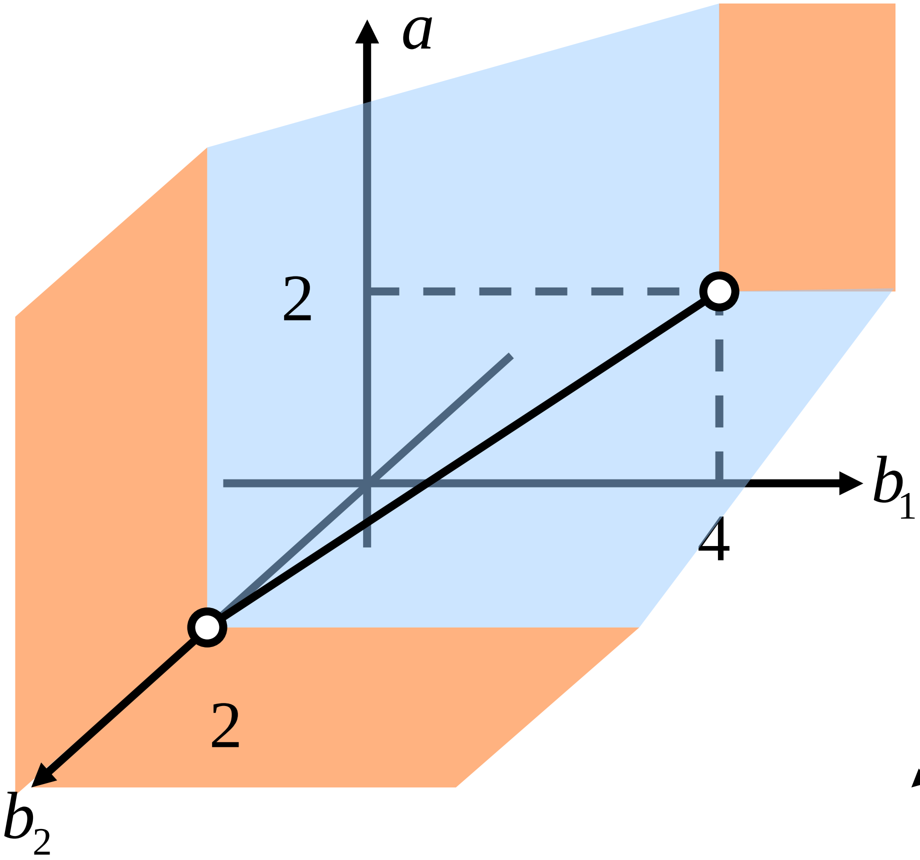
$$= (\underline{ap_1} + (\underline{1} - a)\underline{p_2})^2 + (ap_1^2 + (1-a)p_2^2)^2 + \ldots$$

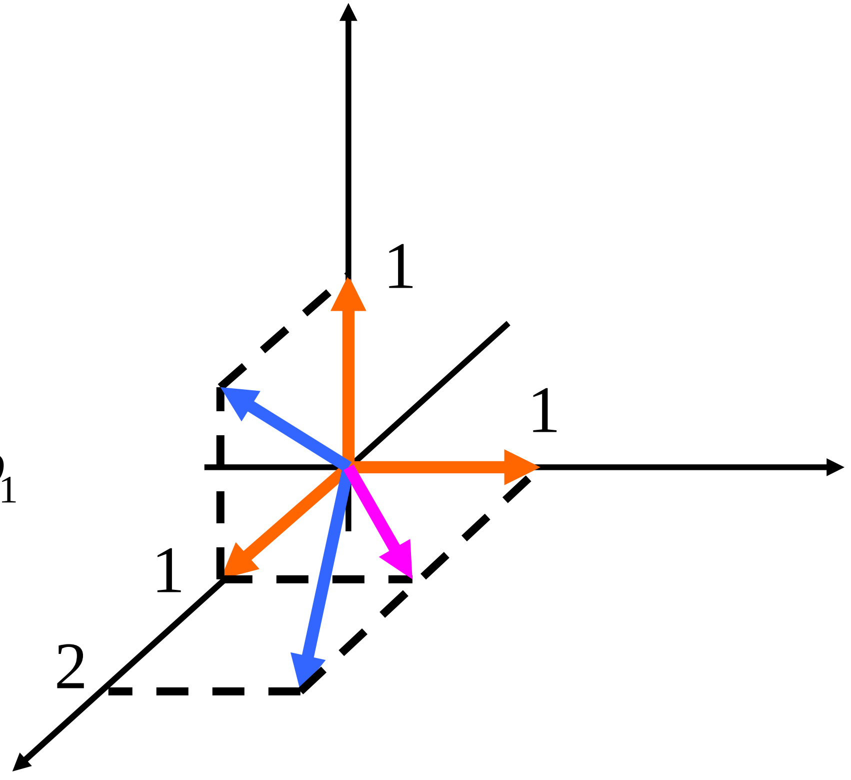
$$= b_2^2 + (ab_1^2 + (b_2 - ab_1)^2)^2 + \ldots$$



The Newton diagram of the mixture

$$H(w) = b_2^2 + (ab_1^2 + (b_2 - ab_1)^2)^2 + \ldots$$



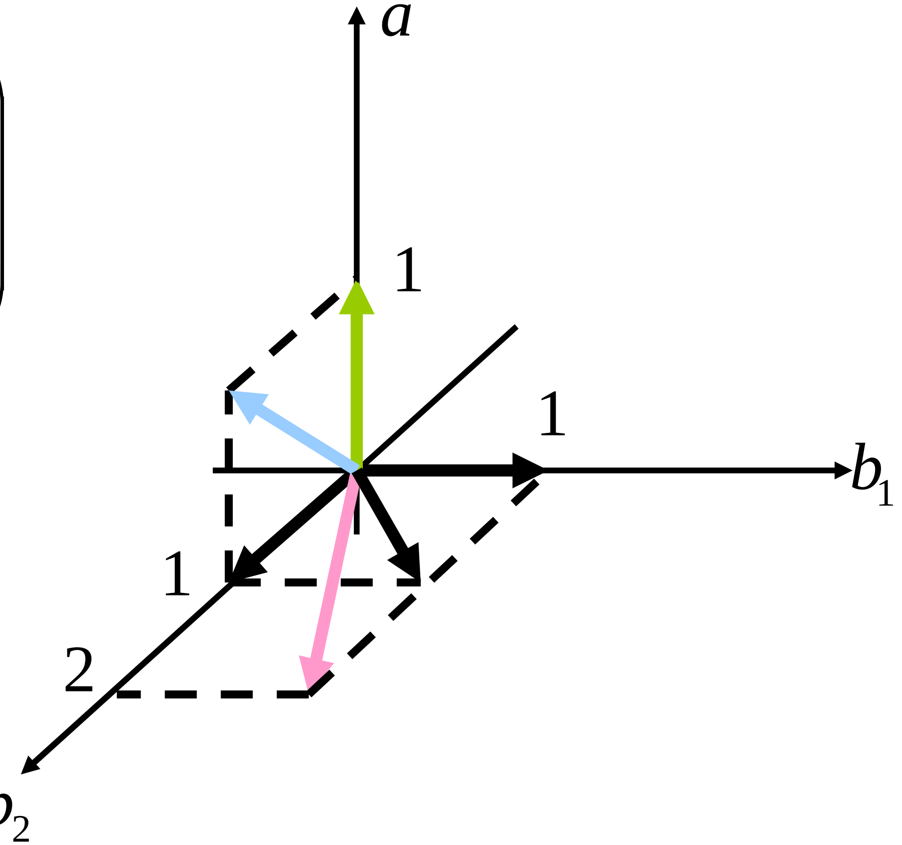


The resolution map based on the toric modification

$$H(w) = b_2^2 + (ab_1^2 + (b_2 - ab_1)^2)^2 + \dots$$

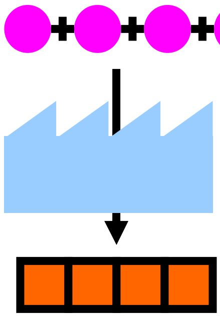
$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 2 & 0 & 1 \end{pmatrix}$$

$$\begin{cases} a = v_1 v_2 \\ b_1 = u \\ b_2 = u^2 v_2 \end{cases}$$



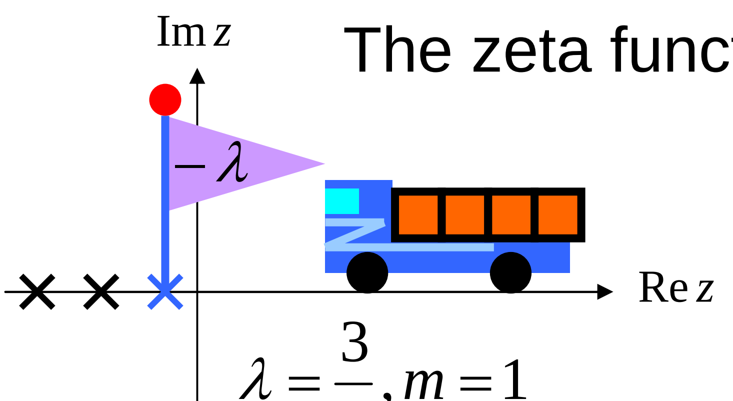
$$H(g(u)) = (1 + v_1^2 v_2^2 + \dots) u^4 v_2^2$$

The generalization error of the mixture of binomial distributions



$$H(w) = b_2^2 + (ab_1^2 + (b_2 - ab_1)^2)^2 + \dots$$: Polynomial form

$$H(g(u)) = (1 + v_1^2 v_2^2 + \dots)u^4 v_2^2$$: Factorized form



$$\lambda = \frac{3}{4}, m = 1$$

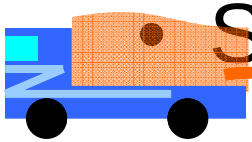
The zeta function:

$$\begin{aligned}
\varsigma(z) &= \int H(g(u))^z \,|\, g'(u) \,|\, du \\
&= \int ((1 + v_1^2 v_2^2 + \dots)u^4 v_2^2)^z \,|\, u^2 v_2 \,|\, du dv_1 dv_2 \\
&= \frac{f(z)}{4z+3} + \dots
\end{aligned}$$

$$G(n) = \frac{3}{4n} + o(1/n \log n)$$: Generalization error

Agenda

- Learning theory and algebraic geometry
- Two forms of the Kullback divergence
- Toric modification
- Application to a binomial mixture model
- Summary



Summary

- The Bayesian generalization error is derived on the basis of the zeta function.
- Calculation of the coefficients requires the factorized form of the Kullback divergence.
- Toric modification is an effective method to find the factorized form.
- The error of a binomial mixture is derived as the application.