Information Theoretic Validity of Penalized Likelihood

Andrew R. Barron

Yale University

Collaborators
Cong Huang, Jonathan Li, Xi Luo, Sabysachi Chatterjee

Cambridge University, StatsLab Seminar, 15 November 2013
Information Theoretic Validity of Penalized Likelihood

Andrew Barron
Outline

1. Some Principles of Information Theory and Statistics
   - Data Compression
   - Minimum Description Length Principle
   - Codelength-Valid Penalties

2. Penalized Likelihood Analysis
   - Redundancy, Resolvability, and Risk
   - Risk-Valid Penalties
   - Penalized Likelihood Analysis for Continuous Parameters
     - $\ell_1$ Penalties are Codelength-Valid and Risk-Valid

3. Regression with $\ell_1$ Penalty
   - Fixed $\sigma^2$ Case
   - Unknown $\sigma^2$ Case

4. Inverse Covariance Estimation with $\ell_1$ type Penalty

5. Summary
Outline

1. Some Principles of Information Theory and Statistics
   - Data Compression
   - Minimum Description Length Principle
   - Codelength-Valid Penalties
2. Penalized Likelihood Analysis
   - Redundancy, Resolvability, and Risk
   - Risk-Valid Penalties
   - Penalized Likelihood Analysis for Continuous Parameters
   - $\ell_1$ Penalties are Codelength-Valid and Risk-Valid
3. Regression with $\ell_1$ Penalty
   - Fixed $\sigma^2$ Case
   - Unknown $\sigma^2$ Case
4. Inverse Covariance Estimation with $\ell_1$ type Penalty
5. Summary
Characterization of uniquely decodeable codelengths

\[ L(x), \quad x \in \mathcal{X}, \quad \sum_x 2^{-L(x)} \leq 1 \]

\[ L(x) = \log \frac{1}{q(x)} \quad q(x) = 2^{-L(x)} \]

Operational meaning of probability:

A distribution is given by a choice of code
Codelessngh Comparison

- Targets $p$ are possible distributions
- Compare codelength $\log \frac{1}{q(x)}$ to targets $\log \frac{1}{p(x)}$
- Redundancy or regret

$$\left[ \log \frac{1}{q(x)} - \log \frac{1}{p(x)} \right]$$

- Expected redundancy

$$D(P_X \parallel Q_X) = E_P \left[ \log \frac{p(X)}{q(X)} \right]$$
Universal Codes

- MODELS
  Family of coding strategies $\iff$ Family of prob. distributions
  Indexed by parameters or functions:

  \[
  \{ L_\theta(x) : \theta \in \Theta \} \iff \{ p_\theta(x) : \theta \in \Theta \}
  \]
  \[
  \{ L_f(x) : f \in \mathcal{F} \} \iff \{ p_f(x) : f \in \mathcal{F} \}
  \]

- Universal codes $\iff$ Universal probabilities $q(x)$

  \[
  L(x) = \log \frac{1}{q(x)}
  \]

- Redundancy: \[
  \left[ \log \frac{1}{q(x)} - \log \frac{1}{p_\theta(x)} \right]
  \]

  Want it small either uniformly in $x, \theta$ or in expectation
Statistical Aim

- Training data $x \Rightarrow$ estimator $\hat{p} = p_{\hat{\theta}}$
- Subsequent data $x'$
- Want $\log 1/\hat{p}(x')$ to compare favorably to $\log 1/p(x')$
- For targets $p$ close to or in the families
Loss

- Kullback Information-divergence:
  \[ D(P_{X'} \parallel Q_{X'}) = E \left[ \log \frac{p(X')}{q(X')} \right] \]

- Bhattacharyya, Hellinger, Chernoff, Rényi divergence:
  \[ d(P_{X'}, Q_{X'}) = 2 \log \frac{1}{E[q(X')/p(X')]^{1/2}} \]

- Product model case: \( p(x') = \prod_{i=1}^{n} p(x'_i) \)
  \[ D(P_{X'} \parallel Q_{X'}) = n D(P \parallel Q) \]
  Likewise \[ d(P_{X'}, Q_{X'}) = n d(P, Q) \]
Loss

- Relationship:
  \[ d(P, Q) \leq D(P \| Q) \]

- and, if the log density ratio is not more than \( B \), then
  \[ D(P \| Q) \leq C_B d(P, Q) \]
  with \( C_B \leq 2 + B \)
Minimum Description Length (MDL) Principle

- Universal coding brought into statistical play

- Minimum Description Length Principle:
  The shortest code for data gives the best statistical model
**MDL: Two-stage Version**

- Two-stage codelength:
  \[
  L(x) = \min_{\theta \in \Theta} \left[ \log \frac{1}{p_\theta(x)} + L(\theta) \right]
  \]
  bits for \( x \) given \( \theta \) + bits for \( \theta \)

- The corresponding statistical estimator is \( \hat{p} = p_{\hat{\theta}} \)

- Typically in \( d \)-dimensional families \( L(\theta) \) is of order
  \[
  \frac{d}{2} \log n
  \]
MDL: Mixture Versions

- Codelength based on a Bayes mixture

\[ L(x) = \log \frac{1}{q(x)} \]

where

\[ q(x) = \int p(x|\theta) w(\theta) d\theta \]

average case optimal and pointwise optimal for a.e. \( \theta \)

- Codelength approximation (Barron 1985, Clarke and Barron 1990, 1994)

\[ \log \frac{1}{p(x|\hat{\theta})} + \frac{d}{2} \log \frac{n}{2\pi} + \log \frac{\hat{I}(\hat{\theta})^{1/2}}{w(\hat{\theta})} \]

where \( \hat{I}(\hat{\theta}) \) is the empirical Fisher Information at the MLE.
Outline

1. Some Principles of Information Theory and Statistics
   - Data Compression
   - Minimum Description Length Principle
   - Code-length-Valid Penalties

2. Penalized Likelihood Analysis
   - Redundancy, Resolvability, and Risk
   - Risk-Valid Penalties
   - Penalized Likelihood Analysis for Continuous Parameters
   - $\ell_1$ Penalties are Code-length-Valid and Risk-Valid

3. Regression with $\ell_1$ Penalty
   - Fixed $\sigma^2$ Case
   - Unknown $\sigma^2$ Case

4. Inverse Covariance Estimation with $\ell_1$ type Penalty

5. Summary

Barron
Information Theoretic Validity of Penalized Likelihood
Two-stage Code Redundancy

- Expected codelength minus target at $p_{\theta^*}$

\[
\text{Redundancy} = E \left[ \min_{\theta \in \Theta} \left\{ \log \frac{1}{p_{\theta}(x)} + L(\theta) \right\} \right] - \log \frac{1}{p_{\theta^*}(x)}
\]

- Redundancy approx in smooth families

\[
\frac{d}{2} \log \frac{n}{2\pi} + \log \left| \frac{I(\theta^*)^{1/2}}{w(\theta^*)} \right|
\]
Redundancy and Resolvability

- Redundancy: $\text{Redundancy} = E \min_{\theta \in \Theta} \left[ \log \frac{p_{\theta^*(x)}}{p_\theta(x)} + L(\theta) \right]$
- Resolvability: $\text{Resolvability} = \min_{\theta \in \Theta} E \left[ \log \frac{p_{\theta^*(x)}}{p_\theta(x)} + L(\theta) \right]$
  $= \min_{\theta \in \Theta} \left[ D(P_{X|\theta^*} \parallel P_{X|\theta}) + L(\theta) \right]$

- Ideal tradeoff of Kullback approximation error & complexity
- Population analogue of the two-stage code MDL criterion
- Divide by $n$ to express as a rate. In the i.i.d. case

$$R_n(\theta^*) = \min_{\theta \in \Theta} \left[ D(\theta^* \parallel \theta) + \frac{L(\theta)}{n} \right]$$
Risk of Estimator based on Two-stage Code

- Estimator $\hat{\theta}$ is the choice achieving the minimization
  \[
  \min_{\theta \in \Theta} \left\{ \log \frac{1}{p_{\theta}(x)} + \mathcal{L}(\theta) \right\}
  \]

- Codelengths for $\theta$ are $\mathcal{L}(\theta) = 2L(\theta)$ with $\sum_{\theta \in \Theta} 2^{-L(\theta)} \leq 1$.

- Total loss $d_n(\theta^*, \hat{\theta})$ with $d_n(\theta^*, \theta) = d(P_{X'|\theta^*}, P_{X'|\theta})$
  \[
  \text{Risk} = E[d_n(\theta^*, \hat{\theta})]
  \]

- Info-Thy bound on risk: (Barron 1985, Barron and Cover 1991, Jonathan Li 1999)
  \[
  \text{Risk} \leq \text{Redundancy} \leq \text{Resolvability}
  \]
Risk of Estimator based on Two-stage Code

- **Risk** $\leq$ Resolvability
- Specialize to i.i.d. case:

$$Ed(\theta^*, \hat{\theta}) \leq \min_{\theta \in \Theta} \left[ D(\theta^* \| \theta) + \frac{L(\theta)}{n} \right]$$

- As $n \nearrow$, tolerate more complex $P_{X|\theta}$ to get near $P_{X|\theta^*}$
- Rate is $1/n$, or close to that rate if the target is simple
- Drawback: Code interpretation entails countable $\Theta$
Key to Risk Analysis

- log likelihood-ratio discrepancy at training $x$ and future $x'$

$$\left[ \log \frac{p_{\theta^*}(x)}{p_\theta(x)} - d_n(\theta^*, \theta) \right]$$

- Proof shows, for $L(\theta)$ satisfying Kraft, that

$$\min_{\theta \in \Theta} \left\{ \left[ \log \frac{p_{\theta^*}(x)}{p_\theta(x)} - d_n(\theta^*, \theta) \right] + L(\theta) \right\}$$

has expectation $\geq 0$. From which the risk bound follows.
Penalized Likelihood for Continuous Parameters

- Penalized Likelihood

$$\min_{\theta \in \Theta} \left\{ \log \frac{1}{p_{\theta}(x)} + \text{Pen}(\theta) \right\}$$

- Possibly uncountable $\Theta$

- Does it have information-theoretic validity?

- Data compression interpretation if there exists a countable $\tilde{\Theta}$ and $L$ satisfying Kraft such that the above is not less than

$$\min_{\tilde{\theta} \in \tilde{\Theta}} \left\{ \log \frac{1}{p_{\tilde{\theta}}(x)} + L(\tilde{\theta}) \right\}$$
Data-Compression Valid Penalties

- Equivalently, $\text{Pen}(\theta)$ is valid for penalized likelihood with uncountable $\Theta$ to have a data compression interpretation if there is such a countable $\tilde{\Theta}$ and Kraft summable $L(\tilde{\theta})$, such that, for every $\theta$ in $\Theta$, there is a representor $\tilde{\theta}$ in $\tilde{\Theta}$ such that

\[
\text{Pen}(\theta) \geq L(\tilde{\theta}) + \log \frac{p_{\theta}(x)}{p_{\tilde{\theta}}(x)}
\]

- This is the link between uncountable and countable cases
Statistical-Risk Valid Penalties

- Penalized Likelihood
  \[ \hat{\theta} = \arg\min_{\theta \in \tilde{\Theta}} \left\{ \log \frac{1}{p_{\theta}(x)} + Pen(\theta) \right\} \]

- Again: possibly uncountable \( \Theta \)
- Is there again a simple resolvability bound on the risk?
- Task: determine a condition on \( Pen(\theta) \) such that the risk is captured by the population analogue

\[ Ed_n(\theta^*, \hat{\theta}) \leq \inf_{\theta \in \Theta} \left\{ E \log \frac{p_{\theta^*}(X)}{p_{\theta}(X)} + Pen(\theta) \right\} \]
Statistical-Risk Valid Penalty

For an uncountable $\Theta$ and a penalty $Pen(\theta)$, $\theta \in \Theta$, suppose there is a countable $\tilde{\Theta}$ and $\mathcal{L}(\tilde{\theta}) = 2\mathcal{L}(\tilde{\theta})$ where $\mathcal{L}(\tilde{\theta})$ satisfies Kraft, such that, for all $x, \theta^*$,

$$\min_{\theta \in \Theta} \left\{ \left( \log \frac{p_{\theta^*}(x)}{p_\theta(x)} - d_n(\theta^*, \theta) \right) + Pen(\theta) \right\} \geq \min_{\tilde{\theta} \in \tilde{\Theta}} \left\{ \log \frac{p_{\theta^*}(x)}{p_{\tilde{\theta}}(x)} - d_n(\theta^*, \tilde{\theta}) + \mathcal{L}(\tilde{\theta}) \right\}$$

The second expression has expectation $\geq 0$, so the first expression does too. Consequently

$$Ed_n(\theta^*, \hat{\theta}) \leq \inf_{\theta \in \Theta} \left\{ E \log \frac{p_{\theta^*}(X)}{p_\theta(X)} + Pen(\theta) \right\}$$

A.B. with J. Li and X. Luo (in Rissanen Festschrift 2008)
Variable Complexity, Variable Distortion Cover

- Equivalent statement of the condition: \( \text{Pen}(\theta) \) is a valid penalty if for each \( \theta \) in \( \Theta \) there is a representor \( \tilde{\theta} \) in \( \tilde{\Theta} \) with complexity \( L(\tilde{\theta}) \), distortion \( \Delta_n(\tilde{\theta}, \theta) \) and

\[
\text{Pen}(\theta) \geq L(\tilde{\theta}) + \Delta_n(\tilde{\theta}, \theta)
\]

where the distortion \( \Delta_n(\tilde{\theta}, \theta) \) is the difference in the discrepancies at \( \tilde{\theta} \) and \( \theta \)

\[
\Delta_n(\tilde{\theta}, \theta) = \log \frac{p_\theta(x)}{p_{\tilde{\theta}}(x)} + d_n(\theta, \theta^*) - d_n(\tilde{\theta}, \theta^*)
\]
A Setting for Regression and log-density Estimation: Linear Span of a Dictionary

- $\mathcal{G}$ is a dictionary of candidate basis functions
  - E.g. wavelets, splines, polynomials, trigonometric terms, sigmoids, explanatory variables and their interactions

- Candidate functions in the linear span
  $$f_\theta(x) = \sum_{g \in \mathcal{G}} \theta_g g(x)$$

- Weighted $\ell_1$ norm of coefficients
  $$\|\theta\|_1 = \sum_{g} a_g |\theta_g|$$

- Weights $a_g = \|g\|_n$ where $\|g\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} g^2(x_i)$
Example Models

- Regression (focus of current presentation)
  \[ p_\theta(y|x) = \text{Normal}(f_\theta(x), \sigma^2) \]

- Logistic regression with \( y \in \{0, 1\} \)
  \[ p_\theta(y|x) = \text{Logistic}(f_\theta(x)) \quad \text{for } y = 1 \]

- Log-density estimation (focus of Festschrift paper)
  \[ p_\theta(x) = \frac{p_0(x) \exp\{f_\theta(x)\}}{c_f} \]

- Gaussian graphical models
\( \ell_1 \) Penalty

- \( \text{pen}(\theta) = \lambda \|\theta\|_1 \) where \( \theta \) are coeff of \( f_\theta(x) = \sum_{g \in G} \theta_g g(x) \)
Outline

1. Some Principles of Information Theory and Statistics
   - Data Compression
   - Minimum Description Length Principle
   - Codelength-Valid Penalties

2. Penalized Likelihood Analysis
   - Redundancy, Resolvability, and Risk
   - Risk-Valid Penalties
   - Penalized Likelihood Analysis for Continuous Parameters
   - $\ell_1$ Penalties are Codelength-Valid and Risk-Valid

3. Regression with $\ell_1$ Penalty
   - Fixed $\sigma^2$ Case
   - Unknown $\sigma^2$ Case

4. Inverse Covariance Estimation with $\ell_1$ type Penalty

5. Summary
Regression with $\ell_1$ penalty

- $\ell_1$ penalized log-density estimation, i.i.d. case
  \[
  \hat{\theta} = \text{argmin}_\theta \left\{ \frac{1}{n} \log \frac{1}{p_{f_\theta}(x)} + \lambda_n \|\theta\|_1 \right\}
  \]

- Regression with Gaussian model, fixed $\sigma^2$
  \[
  \min_{\theta} \left\{ \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\theta(x_i))^2 + \frac{1}{2} \log 2\pi \sigma^2 + \frac{\lambda_n}{\sigma} \|\theta\|_1 \right\}
  \]

- Valid for
  \[
  \lambda_n \geq \sqrt{\frac{2 \log(2M_G)}{n}} \quad \text{with} \quad M_G = \text{Card}(G)
  \]
Adaptive risk bound

- For log density estimation with suitable $\lambda_n$
  \[ Ed(f^*, f_{\hat{\theta}}) \leq \inf_{\theta} \left\{ D(f^* \| f_\theta) + \lambda_n \| \theta \|_1 \right\} \]

- For regression with fixed design points $x_i$, fixed $\sigma$, and $\lambda_n = \sqrt{\frac{2 \log(2M)}{n}}$
  \[ E \| f^* - f_{\hat{\theta}} \|_n^2 \leq \inf_{\theta} \left\{ \frac{\| f^* - f_\theta \|_n^2}{2\sigma^2} + \lambda_n \frac{\| \theta \|_1}{\sigma} \right\} \]
Adaptive risk bound specialized to regression

- Again for fixed design and $\lambda_n = \sqrt{\frac{2\log 2M}{n}}$, multiplying through by $4\sigma^2$,
  \[
  E \| f^* - \hat{f}_\theta \|_n^2 \leq \inf_{\theta} \left\{ 2 \| f^* - f_\theta \|_n^2 + 4\sigma \lambda_n \| \theta \|_1 \right\}
  \]

- In particular for all targets $f^* = f_{\theta^*}$ with finite $\| \theta^* \|$ the risk bound $4\sigma \lambda_n \| \theta^* \|$ is of order $\sqrt{\frac{\log M}{n}}$

Comments on proof

- Likelihood discrepancy plus complexity

\[
\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - f_{\tilde{\theta}}(x_i))^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - f_{\theta}(x_i))^2 + K \log(2M)
\]

- Representor \( f_{\tilde{\theta}} \) of \( f_{\theta} \) of following form, with \( v \) near \( \|\theta\|_1 \)

\[
f_{\tilde{\theta}}(x) = \frac{v}{K} \sum_{k=1}^{K} g_k(x) / \|g_k\|
\]

- \( g_1, \ldots, g_K \) picked at random from \( G \), independently, where \( g \) arises with probability proportional to \( |\theta_g|a_g \)

- Shows exists representor with like. discrep. + complexity

\[
\frac{nv^2}{2K} + K \log(2M)
\]
Comments on proof

- Optimizing it yields penalty proportional to $\ell_1$ norm
- Penalty $\lambda \|\theta\|_1$ is valid for both data compression and statistical risk requirements for $\lambda \geq \lambda_n$ where
  \[
  \lambda_n = \sqrt{\frac{2 \log(2M)}{n}}
  \]
- Especially useful for very large dictionaries
- Improvement for small dictionaries gets rid of log factor: $\log(2M)$ may be replaced by $\log(2e \max\{\frac{M}{\sqrt{n}}, 1\})$
Comments on proof

- Existence of representor shown by random draw is a Shannon-like demonstration of the variable cover (code)
- Similar approximation in analysis of greedy computation of $\ell_1$ penalized least squares
Fixed $\sigma$ versus unknown $\sigma$

- MDL with $\ell_1$ penalty for each possible $\sigma$. Recall

$$\min_{\theta} \left\{ \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\theta(x_i))^2 + \frac{1}{2} \log 2\pi \sigma^2 + \frac{\lambda n}{\sigma} \|\theta\|_1 \right\}$$

- Provides a family of fits indexed by $\sigma$.
- For unknown $\sigma$ suggest optimization over $\sigma$ as well as $\theta$

$$\min_{\theta, \sigma} \left\{ \frac{1}{2\sigma^2} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\theta(x_i))^2 + \frac{1}{2} \log 2\pi \sigma^2 + \frac{\lambda n}{\sigma} \|\theta\|_1 \right\}$$

- Slight modification of this does indeed satisfy our condition for an information-theoretically valid penalty and risk bound (details in the WITMSE 2008 proceedings)
Best $\sigma$

- Best $\sigma$ for each $\theta$ solves the quadratic equation

$$\sigma^2 = \sigma \lambda_n \|\theta\|_1 + \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\theta(x_i))^2$$

- By the quadratic formula the solution is

$$\sigma = \frac{1}{2} \lambda_n \|\theta\|_1 + \sqrt{\left(\frac{1}{2} \lambda_n \|\theta\|_1\right)^2 + \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\theta(x_i))^2}$$
Some Principles of Information Theory and Statistics

Penalized Likelihood Analysis

Regression with ℓ₁ Penalty

Inverse Covariance Estimation with ℓ₁ type Penalty

Outline

1. Some Principles of Information Theory and Statistics
   - Data Compression
   - Minimum Description Length Principle
   - Codelength-Valid Penalties

2. Penalized Likelihood Analysis
   - Redundancy, Resolvability, and Risk
   - Risk-Valid Penalties
   - Penalized Likelihood Analysis for Continuous Parameters
   - ℓ₁ Penalties are Codelength-Valid and Risk-Valid

3. Regression with ℓ₁ Penalty
   - Fixed σ² Case
   - Unknown σ² Case

4. Inverse Covariance Estimation with ℓ₁ type Penalty

5. Summary

Barron

Information Theoretic Validity of Penalized Likelihood
Current work with Sabyasachi Chatterjee

\[ x_1, x_2 \ldots x_n \text{ iid } N(0, \Sigma), \quad \Sigma^{-1} = M. \]

\[ \hat{M} = \arg\min_{M \succ 0, M = M^T} \left\{ \frac{1}{2} (\text{Tr}(MS) - \log \det(M)) + \lambda_n \| M \|_1 \right\} \]

\( \| M \|_1 = \sum_{i \neq j} |M_{ij}| \)

\( S = (1/n) \sum_{i=1}^{n} x_i x_i^T \) is the sample covariance matrix.

\( \hat{M} \) is designed to be sparse.

\( M_{ij} = 0 \) means cond indep of \( X_i \) and \( X_j \) given others.
The Bhattacharya distance between two Gaussians with 0 mean and covariance matricies $\Sigma_1$, $\Sigma_2$ turns out to be

$$\frac{1}{2} \log \left( \frac{\det(\frac{\Sigma_1 + \Sigma_2}{2})}{\sqrt{\det\Sigma_1 \det\Sigma_2}} \right)$$

With similar multiplier $\lambda_n$ the $\ell_1$ penalty is risk valid.
The distortion $\triangle(\tilde{M}, M)$ relative to covariance $\Sigma^*$ is

$$\frac{1}{2} \left( \text{Tr}(\tilde{M}S) - \text{Tr}(MS) \right) - \log \det(I + \tilde{M}\Sigma^*) + \log \det(I + M\Sigma^*)$$

A representer is constructed by a linear combination of matrices of the form $\delta_{ij} = (e_i e_j^T + e_j e_i^T)/2$.

At $k$th step a probabilistic argument is employed to upper bound the minimum distortion of $k$ term representers.
Outline

1. Some Principles of Information Theory and Statistics
   - Data Compression
   - Minimum Description Length Principle
   - Codelength-Valid Penalties

2. Penalized Likelihood Analysis
   - Redundancy, Resolvability, and Risk
   - Risk-Valid Penalties
   - Penalized Likelihood Analysis for Continuous Parameters
   - $\ell_1$ Penalties are Codelength-Valid and Risk-Valid

3. Regression with $\ell_1$ Penalty
   - Fixed $\sigma^2$ Case
   - Unknown $\sigma^2$ Case

4. Inverse Covariance Estimation with $\ell_1$ type Penalty

5. Summary

Barron

Information Theoretic Validity of Penalized Likelihood
Summary

- Allow penalized likelihoods with continuous domains for $\theta$
- **Information-theoretically valid penalties**: Penalty exceed complexity plus distortion of optimized representor of $\theta$
- Yields MDL interpretation and risk controlled by resolvability
- $\ell_0$ penalty $\frac{\text{dim}}{2} \log n$ classically analyzed
- $\ell_1$ penalty $\lambda_n \| \theta \|_1$ analyzed here: Information-theoretically valid for

$$\lambda_n \geq \sqrt{2(\log 2M)/n}$$