Concentration Inequalities and Data-Dependent Error Bounds

Olivier Bousquet
Max Planck Institute for Biological Cybernetics
Tübingen

Jena, 11th February 2003

Overview

- Concentration Inequalities
- Empirical Processes
- Modulus of Continuity
- Data-Dependent Modulus of Continuity
- Statistical Applications

Motivation

Let $X_1, \ldots, X_n$ be n independent random variables

Define

$$Z = f(X_1, \ldots, X_n)\,,$$

Given knowledge about the distribution of the X_i and the function f, what can be said about the distribution of Z ?

We want tail bounds of the form

$$\mathbb{P}\left[Z \geq \mathbb{E}\left[Z\right] + t\right] \leq \delta(t)\,,$$

or with probability at least $1 - \delta$,

$$Z \leq \mathbb{E}\left[Z\right] + B(\delta)\,.$$

Concentration refers to the behavior as a function of n (cf isoperimetry, concentration of Gaussian measure on n-sphere).

Applications

- Sums of independent real-valued random variables

$$Z = \sum X_i \, .$$

- Norms of sums of random vectors in a Banach space

$$Z = \left\| \sum X_i \right\| \, .$$

- Suprema of empirical processes (statistics, learning theory)

$$Z = \sup_{f \in \mathcal{F}} \sum f(X_i) \, .$$

- Functionals of random matrices (e.g. trace, norms...)

$$Z = \|(X_{i,j})\| \, .$$

- Combinatorics, random graphs (e.g. triangles)

$$Z = \sum_{i \neq j \neq k} X_{i,j} X_{j,k} X_{k,i} \, .$$

Sums of real-valued random variables

Let $Z = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Hoeffding's inequality

Theorem 1 (Hoeffding, 1963) *Assume* $X_i \in [0,1]$ *almost surely. Then for all* $x > 0$*, with probability* $1 - e^{-x}$,

$$Z \leq \mathbb{E}[Z] + \sqrt{x/2n}\,.$$

Bennett's inequality

Theorem 2 (Bennett, 1963) *Assume* $\mathbb{E}[X_i] = 0$*,* $X_i \leq 1$ *and* $\sigma^2 = \frac{1}{n}\sum Var[X_i]$*. Then for all* $x > 0$*, with probability* $1 - e^{-x}$,

$$Z \leq \mathbb{E}[Z] + \sqrt{2x\sigma^2/n} + x/3n\,.$$

Concentration inequalities

Recall

$$Z = f(X_1, \ldots, X_n) .$$

Define for all $k = 1, \ldots, n$,

$$Z_k = f_k(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n) .$$

Results on Z are based on conditions on the increments.

$$Z - Z_k$$

McDiarmid's inequality

Theorem 3 (McDiarmid, 1989) *Assume* $n(Z - Z_k) \in [0, 1]$, *then for all* $x > 0$ *with probability at least* $1 - e^{-x}$,

$$Z \leq \mathbb{E}[Z] + \sqrt{x/2n} .$$

Suprema of empirical processes with bounded functions.

Sub-additive functions

Theorem 4 (Boucheron, Lugosi and Massart 2000) *Assume* $n(Z - Z_k) \in [0,1]$ *and* $\sum_{k=1}^n Z - Z_k \leq Z$. *Then for all* $x > 0$, *with probability at least* $1 - e^{-x}$,

$$Z \leq \mathbb{E}[Z] + \sqrt{2x\mathbb{E}[Z]/n} + x/3n\,.$$

Size of the largest subsequence satisfying a certain (hereditary) property.
Suprema of empirical processes with non-negative bounded functions.

Theorem 5 (B. 2002) *Assume* $Y_k \leq n(Z - Z_k) \leq 1$, $\mathbb{E}[Y_k] \geq 0$, $\sigma^2 = \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[Y_k^2\right]$ *and also* $\sum_{k=1}^n Z - Z_k \leq Z$. *Then for all* $x > 0$, *with probability at least* $1 - e^{-x}$,

$$Z \leq \mathbb{E}[Z] + \sqrt{2x(\sigma^2 + 2\mathbb{E}[Z])/n} + x/3n\,.$$

Suprema of empirical processes with upper bounded functions.

Idea of proof

Let ϕ be a convex non-negative function such that $1/\phi''$ is concave. ϕ-entropy

$$H_\phi(Z) = \mathbb{E}\left[\phi(Z)\right] - \phi(\mathbb{E}\left[Z\right]) .$$

Properties

- Non-negative, convex, lower semi-continuous
- Tensorization

$$H_\phi(Z) \leq \mathbb{E}\left[\sum_{k=1..n} H_{\phi,k}(Z)\right] .$$

- $\phi(x) = x^2$ Efron-Stein inequality

$$\mathrm{Var}\left[Z\right] \leq \mathbb{E}\left[\sum_{k=1..n} (Z - Z_k)^2\right] .$$

- $\phi(x) = x \log x$ Modified log-Sobolev inequality (Ledoux, 1996)

$$\mathbb{E}\left[Ze^{\lambda Z}\right] - \mathbb{E}\left[e^{\lambda Z}\right] \log \mathbb{E}\left[e^{\lambda Z}\right] \leq \mathbb{E}\left[\sum_{k=1}^{n} \psi(\lambda(Z - Z_k))e^{\lambda Z}\right] .$$

Empirical Processes

Notation $Pf = \mathbb{E}\left[f(X)\right]$, $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$.

- Let $\mathcal{F}$ be such that $f \in \mathcal{F}$ implies $f(x) \in [0,1]$. McDiarmid's inequality gives

$$\sup_{f\in\mathcal{F}} Pf - P_n f \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right] + \sqrt{2x/n}\,.$$

- Symmetrization

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} Pf - P_n f\right] \leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right]\,.$$

- Consequence

$$\sup_{f\in\mathcal{F}} Pf - P_n f \leq 2\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\right] + \sqrt{8x/n}\,.$$

Empirical Processes

Theorem 6 (B. 2002) *Let $X_i \in \mathcal{X}$ and let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$ such that $f - Pf \leq 1$. Then for all $x > 0$, with probability $1 - e^{-x}$, for all $f \in \mathcal{F}$,*

$$Pf - P_n f \leq \inf_{\alpha > 0} \left((1+\alpha)\mathbb{E}\left[\sup_{f' \in \mathcal{F}} Pf' - P_n f' \right] + \sqrt{2x\sigma^2/n} + (1/3 + 1/\alpha)x/n \right),$$

with $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sup_{f\in\mathcal{F}} \ Var[f(X_i)]$.

How to improve it: $\rightarrow$ Making the right-hand side depend on f

1. restrict the supremum to functions with variance less than $\mathrm{Var}[f]$
2. replace σ^2 by $\mathrm{Var}[f]$

$$\mathrm{Var}[f] \leq r,\ Pf - P_n f \leq c_1 \mathbb{E}\left[\sup_{\substack{f' \in \mathcal{F} \\ \mathrm{Var}[f'] \leq r}} Pf' - P_n f' \right] + c_2\sqrt{xr/n} + c_3 x/n\,.$$

Making this uniform in r ?

Modulus of continuity

- Modulus of continuity at the origin

$$w(\mathcal{F}, r) = \mathbb{E}\left[\sup_{f\in\mathcal{F},\, Pf^2\leq r} |Pf - P_n f|\right] .$$

- We want to have

$$Pf - P_n f \leq c_1 w(\mathcal{F}, Pf^2) + c_2\sqrt{xPf^2/n} + c_3 x/n \,.$$

- Typical behavior of w:

$$w(\mathcal{F}, r) \approx \sqrt{Ar} \,.$$

  Note that A is the solution of $w(\mathcal{F}, r) = r$.

Fixed point

- Sub-root function.
  ϕ non-negative, non-decreasing and $\phi(r)/\sqrt{r}$ is non-increasing.

- Fixed point.
  If there exists ϕ sub-root with
  $$w(\mathcal{F}, r) \leq \phi(r)\,,$$
  then
  $$\phi(r) = r\,,$$
  has a unique solution $r^* > 0$ and we have
  $$w(\mathcal{F}, r) \leq \sqrt{r^* r}\,.$$

Result

Let $\mathcal{F}$ be a class of functions with ranges in $[-1, 1]$

Theorem 7 (B. 2002) *Let r^* be the fixed point of $\phi(r)$. For all $x > 0$ and all $K > 1$, with probability at least $1 - e^{-x}$*

$$|Pf - P_n f| \leq K^{-1} P f^2 + cKr^* + c' K \frac{x}{n}\,.$$

More generally if $\kappa \geq 1$,

$$|Pf - P_n f| \leq K^{-1} (P f^2)^\kappa + cK^{2\gamma-1} (r^*)^\gamma + c' K^{2\gamma-1} (\frac{x}{n})^\gamma\,.$$

with $\gamma = \kappa/(2\kappa - 1)$.

$\longrightarrow$ Further improvement ? Computing r^* from the data ?

Data-dependent modulus of continuity

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F},\, P_n f^2 \leq r} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right] \leq \phi_n(r)\,.$$

Theorem 8 (B. 2002) *Let r_n^* be the fixed point of $\phi_n(r)$. For all $x > 0$ and all $K > 1$, with probability at least $1 - e^{-x}$*

$$|Pf - P_n f| \leq K^{-1} P f^2 + cK r_n^* + c' K \frac{x + \log\log n}{n}\,.$$

The Learning Problem

Problem: Learning from examples

- Observe a set of objects (inputs) $X_1, \ldots, X_n$ with their associated label (output) $Y_1, \ldots, Y_n$.
- Goal: for a new, unobserved object X, predict Y.

Formalization

- $(X, Y) \sim P$ pair of random variables, values in $\mathcal{X} \times \mathcal{Y}$, P unknown joint distribution.
- Given n i.i.d. pairs (X_i, Y_i) sampled according to P, find $g : \mathcal{X} \to \mathcal{Y}$ such that $P(g(X) \neq Y)$ is small

More generally, ℓ measures the cost of errors. Minimize

$$L(g) = \mathbb{E}\left[\ell(g(X), Y)\right]$$

Possible Algorithms

Goal: minimize $L(g) = \mathbb{E}\left[\ell(g(X), Y)\right]$.

- Empirical risk minimization (ERM): approximate the risk by $L_n(g) = \frac{1}{n}\sum_{i=1}^n \ell(g(X_i), Y_i)$ and solve

$$\min_{g \in \mathcal{G}} L_n(g)\,.$$

- Structural risk minimization (SRM)/Model selection: several 'models' $\{\mathcal{G}_m : m \in \mathcal{M}\}$ and solve

$$\min_{m \in \mathcal{M}} \min_{g \in \mathcal{G}_m} L_n(g) + p(m)\,.$$

- Regularization: introduce a weight functional $w(g)$ and solve

$$\min_{g \in \mathcal{G}} L_n(g) + \lambda w(g)\,.$$

This covers most algorithms (SVM, Boosting...).

Application to estimation

$$\mathbb{E}\left[\sup_{g,g'\in\mathcal{G}:P(g-g')^2\le r} |\frac{1}{n}\sum_{i=1}^{n}\eta_i(g(X_i)-g'(X_i))|\right] \le \phi(r)\,.$$

Corollary 1 *Let $\mathcal{G}$ be a class of functions such that*
$\mathbb{E}\left[(\ell_g-\ell_s)^2\right] \le (L(g)-L(s))^{1/\kappa}$.
Then with probability $1-e^{-x}$,

$$L(g)-L(s) \le c\left(L(g^*)-L(s)+(r^*)^{\kappa/(2\kappa-1)}+(x/n)^{\kappa/(2\kappa-1)}\right)\,.$$

- Assumption satisfied if noise benign (Tsybakov).
- Minimax rates under Tsybakov's conditions for VC classes
- Fixed point of modulus of continuity as a measure of the complexity
- Modulus on the initial class (Gaussian contraction)

Data-dependent error bounds

$$\mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}:P_n(g-g_n)^2\leq r}\frac{1}{n}\sum_{i=1}^n\sigma_i(g(X_i)-g_n(X_i))\right]\leq\phi_n(r)\,.$$

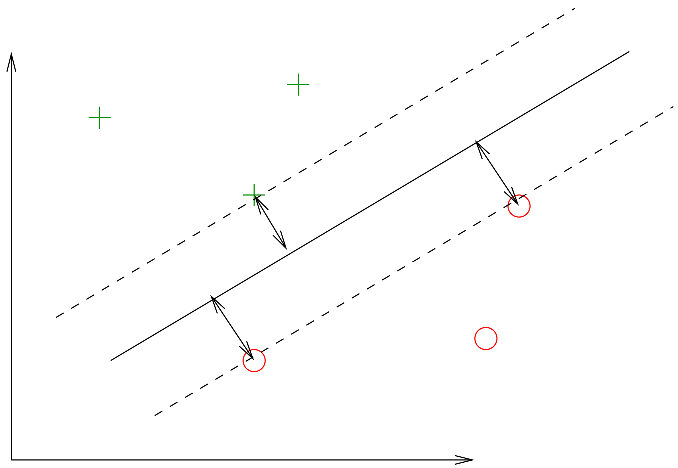
- Conditional process (data is fixed)
- Computed at the empirical error minimizer g_n

Theorem 9 (B. 2002) *Let $\mathcal{G}$ be a class of functions such that* $\mathbb{E}\left[(\ell_g-\ell_s)^2\right]\leq L(g)-L(s)$.
Let r_n^* be the fixed point of ϕ_n. Then with probability $1-e^{-x}$,

$$L(g)-L(s)\leq c\left(L(g^*)-L(s)+r_n^*+(x+\log\log n)/n\right)\,.$$

$\longrightarrow$ r_n^* can be computed from the data only.

Application to SVM



Consider $Y \in \{-1, 1\}$. The SVM algorithm solves

$$\min_{g \in \mathcal{G}_k} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i g(X_i))_+ + \lambda \left\| g \right\|^2 ,$$

in a reproducing kernel Hilbert space $\mathcal{G}_k$ generated by $k(x, x')$.

- Properties of the loss (with benign noise)
- Modulus of continuity ?

Application to SVM

Properties of the loss

Regression function: $s(x) = \mathbb{P}\left[Y = 1 \mid X = x\right]$ $(L(s) = \inf L)$

Bayes classifier: $\eta^*(x) = 1$ if $s(x) > 1/2$ and -1 otherwise

$$L(\eta^*) = L(s)\,.$$

Lemma 1 *For any function g,*

$$\mathbb{P}\left[Y g(X) \leq 0\right] - \mathbb{P}\left[Y \eta^*(X) \leq 0\right] \leq L(g) - L(\eta^*)\,.$$

$\longrightarrow$ Difference in misclassification error bouded by difference in loss

Lemma 2 *Assume that $|s(X) - 1/2| \geq \eta_0$ a.s. If $\|g\|_\infty \leq M$ then*

$$\mathbb{E}\left[(\ell(g) - \ell(\eta^*))^2\right] \leq \left(M - 1 + \eta_0^{-1}\right)(L(g) - L(\eta^*))\,.$$

$\longrightarrow$ If noise is nice, variance linearly related to expectation

Application to SVM

Capacity Bound

Gram matrix from the data $K = (k(X_i, X_j))_{i,j}$

Eigenvalues of K, $\lambda_1 \geq \lambda_2 \geq \ldots$

Space of functions ellipsoid shaped (eigenvalues)

- Volume-based (covering numbers) $\prod_{i \geq 1} \lambda_i$
- Rademacher $\sqrt{\sum_{i \geq 1} \lambda_i / n}$

Theorem 10 (B. 2002)

$$r_n^* \leq \frac{c}{n} \inf_{d \in \mathbb{N}} \left(d + \sqrt{\sum_{j > d} \lambda_j} \right) .$$

- Trace corresponds to $d = 0$
- Exponential decay (RBF kernel) gives $\log n / n$ instead of $1/\sqrt{n}$
- Data-dependent, explicit constants

Application to Boosting

Space of functions $\mathcal{F}$

$$\min_{g \in \mathrm{conv}(\mathcal{F})} \frac{1}{n} \sum_{i=1}^{n} e^{-Y_i g(X_i)} + \lambda \left\| g \right\|_1 .$$

Loss: treated by Lugosi and Vayatis

Capacity: ω modulus of continuity of conditional Gaussian process

Theorem 11 (B., Koltchinskii and Panchenko 2002)

$$\omega(\mathrm{conv}(\mathcal{F}), r) \leq \inf_{\epsilon} \left(2\omega(\mathcal{F}, r) + r\sqrt{N(\mathcal{F}, \epsilon)} \right) ,$$

where N is the covering number.

Conclusion

1. Data-dependent bounds

2. involving modulus of continuity of Rademacher conditional process

3. computed on the initial class $\mathcal{G}$

4. minimax rates under various conditions

$\longrightarrow$ New quantities involved in the bounds

$\longrightarrow$ New algorithms