Bias reduction and efficiency in estimation of smooth functionals of high-dimensional parameters

Vladimir Koltchinskii

School of Mathematics
Georgia Institute of Technology

HDP-2020

Estimation of functionals

- $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta$
- $\Theta \subset E$, E a linear normed space
- $f : \Theta \mapsto \mathbb{R}$ a functional (that possesses some "smoothness")
- Problems:
  - What is the optimal mean squared error rate of estimation of $f(\theta)$ based on $X_1, \dots, X_n$?
  - Is there a threshold on the degree of smoothness s of functional f such that estimation of $f(\theta)$ with $\sqrt{n}$-rate is possible for s above the threshold?
  - For which degree of smoothness efficient estimation of $f(\theta)$ with $\sqrt{n}$-rate is possible?

Estimation of functionals: some references

- Levit (1975, 1978)
- Ibragimov, Nemirovski and Khasminski (1987)
- Bickel and Ritov (1988)
- Donoho and Nussbaum (1990)
- Nemirovski (1990, 2000)
- Birgé and Massart (1995)
- Laurent (1996)
- Lepski, Nemirovski and Spokoiny (1999)
- Cai and Low (2005)
- Robins, Li, Tchetgen and van der Vaart (2008)
- Collier, Comminges and Tsybakov (2017)

Estimation of functionals: some references

- Ibragimov, Nemirovski and Khasminskii (1987), Nemirovski (1990, 2000) studied estimation of smooth functionals of parameter of infinite-dimensional Gaussian shift model and showed that there exists a “smoothness threshold" such that the efficient estimation with parametric convergence rate $\sqrt{n}$ is possible when the smoothness of the functional is above the threshold and is impossible (for some functionals) otherwise. The smoothness threshold depends on the rate of decay of Kolmogorov diameters of the parameter set.

Main Results for Normal Model (K (2017, 2018), K& Zhilova (2018), K& Zhilova (2019))

- $X_1, \dots, X_n$ i.i.d. $N(\mu, \Sigma)$ in $\mathbb{R}^d$ (equipped with the Euclidean norm)
- $\theta := (\mu, \Sigma)$ unknown parameter
- $\Theta := \mathbb{R}^d \times \mathcal{C}_+^d$ the parameter space
- $\mathcal{S}^d$ the space of symmetric $d \times d$ matrices (equipped with the operator norm)
- $\|(w, W)\| := \|w\| + \|W\|, (w, W) \in \mathbb{R}^d \times \mathcal{S}^d$
- $\mathcal{C}_+^d \subset \mathcal{S}^d$ the cone of covariance matrices in $\mathbb{R}^d$
- $\hat{\theta} := (\hat{\mu}, \hat{\Sigma})$,

$$\hat{\mu} = \bar{X} = \frac{X_1 + \cdots + X_n}{n}, \quad \hat{\Sigma} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X}) \otimes (X_j - \bar{X})$$

- $$\mathbb{E}\|\hat{\theta} - \theta\| \lesssim \|\Sigma\|^{1/2} \sqrt{\frac{d}{n}} + \|\Sigma\| \left(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \right).$$

Smoothness

- $g : \Theta \mapsto \mathbb{R}$ k times Frèchet differentiable for $k \geq 0$
- $$\|g\|_{C^k} := \max_{0 \leq j \leq k} \sup_{\theta \in \Theta} \|g^{(j)}(\theta)\|$$
- For $s = k + \rho$ with $k \geq 0$ and $\rho \in (0, 1)$, define
$$\|g\|_{C^s} := \|g\|_{C^k} + \sup_{\theta, \theta' \in \Theta, \theta \neq \theta'} \frac{\|g^{(k)}(\theta) - g^{(k)}(\theta')\|}{\|\theta - \theta'\|^\rho}.$$
- Note: norms of the derivatives are defined as the operator norms (of multilinear forms).
- $C^s(\Theta) := \{g : \Theta \mapsto \mathbb{R} : \|g\|_{C^s(\Theta)} < \infty\}$

Loss functions and Orlicz norm

- $\ell : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a convex nondecreasing loss function such that $\ell(0) = 0$
- $\|\xi\|_\ell := \|\xi\|_{L_\ell(\mathbb{P})} := \inf\Big\{c > 0 : \mathbb{E}\ell\Big(\frac{|\xi|}{c}\Big) \leq 1\Big\}$
- For $\ell(u) := u^p, p \geq 1$, $\|\xi\|_\ell = \|\xi\|_{L_p}$
- Other choices: $\ell(u) = \psi_1(u) := e^u - 1$ (subexponential loss) and $\ell(u) = \psi_2(u) = e^{u^2} - 1$ (subgaussian loss)

Risk Bounds

Let

$$\Theta(a;d) := \mathbb{R}^d \times \left\{\Sigma \in \mathcal{C}_+^d : \sigma(\Sigma) \subset [1/a, a]\right\}, a \geq 1,$$

$\sigma(\Sigma)$ being the spectrum of Σ

Theorem

Suppose that $f \in C^s(\Theta)$ for some $s = k + 1 + \rho, k \geq 0, \rho \in (0,1]$ and that $\ell(u) \leq e^{bu}, u \geq 0$ for some $b > 0$. Then, there exists a functional $f_k : \Theta \mapsto \mathbb{R}$ such that

$$\sup_{\theta \in \Theta(a;d)} \|f_k(\hat{\theta}) - f(\theta)\|_{L_\ell(\mathbb{P}_\theta)} \lesssim_{s,\ell,a} \|f\|_{C^s} \left[\left(\frac{1}{\sqrt{n}} \bigvee \left(\sqrt{\frac{d}{n}}\right)^s\right) \wedge 1\right].$$

Risk Bounds

Theorem

The following minimax bound holds:

$$\sup_{\|f\|_{C^s}\leq 1} \inf_T \sup_{\theta\in\Theta(a;d)} \|T - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \asymp_{a,s} \left(\left(\frac{1}{\sqrt{n}} \bigvee \left(\sqrt{\frac{d}{n}}\right)^s\right) \wedge 1\right),$$

where the infimum is taken over all estimators $T = T(X_1, \dots, X_n)$

Risk Bounds

- If $d = d_n \leq n^{\alpha}$ for some $\alpha \in (0, 1)$ and $s \geq \frac{1}{1-\alpha}$, then

$$\sup_{\theta \in \Theta(a;d)} \|f_k(\hat{\theta}) - f(\theta)\|_{L_\ell(\mathbb{P}_\theta)} = O(n^{-1/2}).$$

- For $d = d_n \geq n^{\alpha}, \alpha \in (0, 1)$ and $s < \frac{1}{1-\alpha}$, there are functionals f with $\|f\|_{C^s} \leq 1$ that could not be estimated with a rate better than $n^{-s(1-\alpha)/2}$, which is slower than $n^{-1/2}$.

Efficiency

Let

$$\sigma_f^2(\theta) := \|\Sigma^{1/2} f'_\mu(\mu,\Sigma)\|^2 + 2\|\Sigma^{1/2} f'_\Sigma(\mu,\Sigma)\Sigma^{1/2}\|_2^2, \theta = (\mu,\Sigma) \in \Theta.$$

For r.v. η_1, η_2 in $\mathbb{R}$, denote
$d_K(\eta_1,\eta_2) := \sup_{x\in\mathbb{R}} |\mathbb{P}\{\eta_1 \le x\} - \mathbb{P}\{\eta_2 \le x\}|$.

Theorem

Suppose $d = d_n \le n^\alpha$ *for some* $\alpha \in (0,1)$. *Then, for all* $s = k + 1 + \rho > \frac{1}{1-\alpha}$, $k \ge 0$, $\rho \in (0,1]$,

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta\in\Theta(a;d_n)} \left| n\mathbb{E}_\theta (f_k(\hat\theta) - f(\theta))^2 - \sigma_f^2(\theta) \right| \to 0, \ n \to \infty$$

and, for all $\sigma_0 > 0$ *and for* $Z \sim N(0,1)$,

$$\sup_{\|f\|_{C^s}\le 1} \sup_{\theta\in\Theta(a;d_n),\sigma_f(\theta)\ge\sigma_0} d_K\left(\frac{\sqrt{n}(f_k(\hat\theta) - f(\theta))}{\sigma_f(\theta)}; Z\right) \to 0, n \to \infty.$$

Efficiency: Minimax Lower Bound

Theorem

Let $f : \Theta \mapsto \mathbb{R}$ *be differentiable with derivative* f' *and let*

$$\omega_{f'}(\theta_0;\delta) := \sup_{\|\theta-\theta_0\|\leq\delta} \|f'(\theta) - f'(\theta_0)\|.$$

Then, for all $\delta > 0$ *and all* $\theta_0 \in \Theta(a;d)$ *such that* $\{\theta : \|\theta - \theta_0\| \leq \delta\} \subset \Theta(a;d)$, *the following bound holds:*

$$\inf_{T_n} \sup_{\|\theta-\theta_0\|\leq\delta} \frac{n\mathbb{E}_\theta\Bigl(T_n(X_1,\dots,X_n) - f(\theta)\Bigr)^2}{\sigma_f^2(\theta)} \geq 1 - D_\beta\left[\frac{a\,\omega_{f'}(\theta_0;\delta)}{\sigma_f(\theta_0)} + a^\beta\delta + \frac{a^2}{\delta^2 n}\right],$$

where the infimum is taken over all estimators $T_n(X_1,\dots,X_n)$.

Bias Reduction in Functional Estimation

- $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta$, $\Theta \subset E$
- $\hat{\theta} \in \Theta$ an estimator of θ based on $X_1, \dots, X_n$
- $\sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\| \lesssim \sqrt{\frac{d}{n}}$
- d is “dimension" (or “complexity") of Θ
- Problem: given a smooth functional $f : \Theta \mapsto \mathbb{R}$, find a functional $g : \Theta \mapsto \mathbb{R}$ such that

$$\mathbb{E}_\theta g(\hat{\theta}) - f(\theta) = O(n^{-1/2}).$$

- We also want

$$g(\hat{\theta}) - \mathbb{E}_\theta g(\hat{\theta}) = O_{\mathbb{P}}(n^{-1/2})$$

that would lead to $\sqrt{n}$ convergence rate.

Operator $\mathcal{T}$

- $$\mathcal{T}g(\theta) := \mathbb{E}_\theta g(\hat{\theta}) = \int_\Theta g(t)P(\theta; dt), \theta \in \Theta,$$

  where

  $$P(\theta; A) := \mathbb{P}_\theta\{\hat{\theta} \in A\}, A \subset \Theta$$

  is a Markov kernel.
- Want to find an approximate solution g of the integral equation $\mathcal{T}g = f$ such that

  $$\mathcal{T}g(\theta) = f(\theta) + O(n^{-1/2}), \theta \in \Theta.$$

Operator $\mathcal{T}$: Gaussian model with unknown mean

- Example. Let $X_1, \dots, X_n \sim$ i.i.d. $N(\theta; I_d), \theta \in \mathbb{R}^d$
- $\hat{\theta} := \bar{X}$
- Heat semigroup.

$$u(\theta, t) = (H^t g)(\theta) = \mathbb{E} g(\theta + w(t))$$
$$= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}^d} g(y) \exp\Big\{-\frac{|\theta - y|^2}{2t}\Big\} dy$$

- Heat equation.

$$\frac{\partial u(\theta, t)}{\partial t} = \frac{1}{2} \Delta u(\theta, t), \ u(\theta, 0) = g(\theta)$$

- $\mathcal{T} g(\theta) = \mathbb{E}_\theta g(\bar{X}) = (H^{1/n} g)(\theta), \theta \in \mathbb{R}^d$
- Solving $\mathcal{T} g = f$ is equivalent to solving an inverse problem for the heat equation (Kolmogorov (1950)).

Operator $\mathcal{T}$: Gaussian model with unknown mean

For a polynomial

$$f(\theta) = \sum_{k_1,\dots,k_d} c_{k_1,\dots,k_d} \theta_1^{k_1} \dots \theta_d^{k_d},$$

the solution g of the equation $\mathcal{T} g = f$ is given by

$$g(x) = \sum_{k_1,\dots,k_d} \frac{c_{k_1,\dots,k_d}}{n^{(k_1+\dots+k_d)/2}} H_{k_1}(\sqrt{n}x_1) \dots H_{k_d}(\sqrt{n}x_d),$$
$$x = (x_1, \dots, x_d) \in \mathbb{R}^d,$$

where $\{H_k\}$ are Hermite polynomials.

Operator $\mathcal{T}$: Gaussian model with unknown covariance

- $X_1, \dots, X_n$ i.i.d. $\sim N(0; \Sigma)$ in $\mathbb{R}^d$, $\Sigma \in \mathcal{C}_+^d$
- $\hat{\Sigma}$ the sample covariance based on $X_1, \dots, X_n$
- Wishart operator:

$$\mathcal{T} g(\Sigma) := \mathbb{E}_\Sigma g(\hat{\Sigma}) = \int_{\mathcal{C}_+^d} g(V) P(\Sigma; dV), \Sigma \in \mathcal{C}_+^d,$$

where Markov kernel $P(\Sigma; \cdot)$ is a rescaled Wishart distribution $\mathcal{W}_d(\Sigma; n)$:

$$P(\Sigma; A) := \mathbb{P}_\Sigma\{\hat{\Sigma} \in A\}, A \subset \mathcal{C}_+^d.$$

Operator $\mathcal{T}$: Gaussian model with unknown covariance

Orthogonally invariant functions g on the cone $\mathcal{C}_+^d$, i.e. functions g such that

$$g(U^{-1}\Sigma U) = g(\Sigma), \Sigma \in \mathcal{C}_+^d$$

for all orthogonal transformations U of $\mathbb{R}^d$, form an invariant subspace of Wishart operator $\mathcal{T}$. Its eigenfunctions in this subspace are zonal polynomials. For an orthogonally invariant polynomial f on $\mathcal{C}_+^d$, the solution of the equation $\mathcal{T}g = f$ could be written in terms of zonal polynomials.

Operators $\mathcal{T}$ and $\mathcal{B}$

- Let $\mathcal{B} := \mathcal{T} - \mathcal{I}$. Informally, $\mathcal{T} g = f$ implies ("Neumann series")

$$g = (\mathcal{I} + \mathcal{B})^{-1} f = (\mathcal{I} - \mathcal{B} + \mathcal{B}^2 - \dots) f.$$

- Define

$$f_k(\theta) := \sum_{j=0}^{k} (-1)^j \mathcal{B}^j f(\theta) = f(\theta) + \sum_{j=1}^{k} (-1)^j \mathcal{B}^j f(\theta), \theta \in \Theta.$$

Then, the bias of estimator $f_k(\hat{\theta})$ is

$$\mathbb{E}_\theta f_k(\hat{\theta}) - f(\theta) = (-1)^k \mathcal{B}^{k+1} f(\theta), \theta \in \Theta.$$

Statistical interpretation of estimator $f_k(\hat{\theta})$

- The bias of plug-in estimator $f(\hat{\theta})$ is
$$\mathbb{E}_\theta f(\hat{\theta}) - f(\theta) = \mathcal{B}f(\theta).$$
- The first order bias correction yields an estimator
$$f(\hat{\theta}) - \mathcal{B}f(\hat{\theta}).$$
- The bias of estimator $\mathcal{B}f(\hat{\theta})$ of $\mathcal{B}f(\theta)$ is
$$\mathbb{E}_\theta \mathcal{B}f(\hat{\theta}) - \mathcal{B}f(\theta) = \mathcal{B}^2 f(\theta).$$
- The second order bias correction yields an estimator
$$f(\hat{\theta}) - \mathcal{B}f(\hat{\theta}) + \mathcal{B}^2 f(\hat{\theta}), \dots$$

- Let
$$\hat{\theta}^{(0)} = \theta \to \hat{\theta}^{(1)} = \hat{\theta} \to \hat{\theta}^{(2)} \to \dots$$
be the Markov chain starting at $\theta \in \Theta$ with transition probability kernel $P(\theta; A), \theta \in \Theta, A \subset \Theta$.
- $\mathcal{T}^k f(\theta) = \mathbb{E}_\theta f(\hat{\theta}^{(k)}), k \geq 0.$
- $$\mathcal{B}^k f(\theta) = (\mathcal{T} - \mathcal{I})^k f(\theta) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \mathcal{T}^j f(\theta)$$
$$= \mathbb{E}_\theta \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(\hat{\theta}^{(j)}).$$

Bootstrap Chain

- Thus, $\mathcal{B}^k f(\theta)$ is the expectation of the k-th order difference of f along the Bootstrap Chain
- Note that $\mathbb{E}_{\hat{\theta}^{(k)}} \|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}\| \lesssim \sqrt{\frac{d}{n}}$.
- Question. Suppose $d \lesssim n$. Is it true that, for $f \in C^k$,

$$|\mathcal{B}^k f(\theta)| \lesssim \left(\sqrt{\frac{d}{n}}\right)^k ?$$

Random Homotopies and Representation of Bootstrap Chain (K& Zhilova, 2019)

Definition

A stochastic process $H(\theta; t), \theta \in \Theta, t \in [0, 1]$ with values in Θ will be called a random homotopy between θ and $\hat{\theta}$ iff

(i) $H(\theta; 0) = \theta, \theta \in \Theta$

(ii) $H(\theta; 1) \stackrel{d}{=} \hat{\theta}(X_1, \dots, X_n), X_1, \dots, X_n \text{ i.i.d. } \sim P_\theta, \theta \in \Theta.$

In addition, some smoothness assumptions on $H(\theta; t), \theta \in \Theta, t \in [0, 1]$ are needed.

Examples of random homotopies

Example 1.

- $X_1, \dots, X_n$ i.i.d $\sim N(\theta; \Sigma), \theta \in E$ unknown mean, E a linear normed space, $\Sigma : E^* \mapsto E$ known covariance operator.
- $\hat{\theta} = \bar{X}$
- $H(\theta; t) = \theta + \frac{t\xi}{\sqrt{n}}, \theta \in E, t \in [0, 1], \xi \sim N(0; \Sigma)$
- $\frac{d}{dt} H(\theta; t) \overset{d}{=} \hat{\theta} - \theta$

Examples of random homotopies

Example 2.

- $X_1, \ldots, X_n$ i.i.d. $\sim N(\mu, \Sigma)$, $\theta = (\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{C}_+^d$.
- $\hat{\theta} = (\hat{\mu}, \hat{\Sigma}), \hat{\mu} := \bar{X}, \hat{\Sigma} := \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}) \otimes (X_j - \bar{X})$
- Let $Z_1, \ldots, Z_n$ be i.i.d. $\sim N(0; I_d)$

$$H(\theta; t) := (1-t)(\mu, \Sigma) + t(\mu + \Sigma^{1/2}\hat{\mu}_Z, \Sigma^{1/2}\hat{\Sigma}_Z\Sigma^{1/2}).$$

- $\frac{d}{dt} H(\theta; t) \overset{d}{=} \hat{\theta} - \theta$

Examples of random homotopies (Moser coupling)

Example 3.

- S compact Riemannian manifold, P normalized Riemannian volume
- $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta$, $P_\theta(dx) = p_\theta(x)P(dx)$, $\Theta \subset E$ a convex set,
- $(x, \theta) \mapsto p_\theta(x)$ smooth, p_θ bounded away from 0
- Moser coupling: Let u_θ be a solution of Poisson equation: $\Delta u_\theta = 1 - p_\theta$ and let

$$v_\theta(t; x) := \frac{\nabla u_\theta(x)}{1 - t + tp_\theta(x)}, x \in S, t \in [0, 1]$$

- The vector field v_θ generates a flow $T_0^t(x)$ on the manifold. Let $g_\theta(x) = T_0^1(x), x \in S$
- Then (Moser): $P_\theta = P \circ g_\theta^{-1}$

Examples of random homotopies (Moser coupling)

Example 3 (cont.)

- Let $U_1, \dots, U_n$ i.i.d. $\sim P$
- $(X_1, \dots, X_n) \stackrel{d}{=} (g_\theta(U_1), \dots, g_\theta(U_n))$
- $\hat{\theta}(X_1, \dots, X_n) \stackrel{d}{=} \hat{\theta}(g_\theta(U_1), \dots, g_\theta(U_n))$
- Let

$$H(\theta; t) := (1-t)\theta + t\hat{\theta}(g_\theta(U_1), \dots, g_\theta(U_n)), \theta \in \Theta, t \in [0,1]$$

- $\frac{d}{dt}H(\theta; t) \stackrel{d}{=} \hat{\theta} - \theta$
- A similar construction could be used when S is not compact and is equipped with probability measure $P(dx) = e^{-V(x)}dx$.

Random Homotopies and Representation of Bootstrap Chain

- $F_1 : \Theta \times [0,1]^k \mapsto \Theta, F_2 : \Theta \times [0,1]^l \mapsto \Theta$
- $F_2 \bullet F_1 : \Theta \times [0,1]^{k+l} \mapsto \Theta,$

$$(F_2 \bullet F_1)(\theta, (t,s)) := F_2(F_1(\theta; t); s), t \in [0,1]^k, s \in [0,1]^l.$$

- Let H be a random homotopy between θ and $\hat{\theta}$ and $H_1, \dots, H_k$ be its i.i.d. copies. Define

$$G_k := H_k \bullet \dots \bullet H_1 : \Theta \times [0,1]^k \mapsto \Theta.$$

Random Homotopies and Representation of Bootstrap Chain

Proposition

(i) Let $\tilde{\theta}_k := G_k(\theta; 1, \ldots, 1), k \geq 1$ and $\tilde{\theta}_0 = \theta$. Then

$$\{\tilde{\theta}_k : k \geq 0\} \overset{d}{=} \{\hat{\theta}_k : k \geq 0\}.$$

(ii) Moreover, for $j \leq k$,

$$\hat{\theta}_j \overset{d}{=} G_k(\theta; t_1, \ldots, t_k),\ (t_1, \ldots, t_k) \in \{0, 1\}^k, \sum_{i=1}^{k} t_i = j.$$

Random Homotopies and Representation of Operators $\mathcal{T}^k, \mathcal{B}^k$

It follows that

$$\mathcal{T}^k f(\theta) = \mathbb{E}_\theta f(\hat{\theta}^{(k)}) = \mathbb{E} f(G_k(\theta; 1, \dots, 1))$$

and, for all $1 \leq j \leq k$, $(t_1, \dots, t_k) \in \{0, 1\}^k$ with $\sum_{i=1}^k t_i = j$,

$$\mathcal{T}^j f(\theta) = \mathbb{E} f(G_k(\theta; t_1, \dots, t_k)).$$

Random Homotopies and Representation of Operators $\mathcal{T}^k, \mathcal{B}^k$

Therefore

$$\mathcal{B}^k f(\theta) = (\mathcal{T} - \mathcal{I})^k f(\theta) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} T^j f(\theta)$$

$$= \sum_{j=0}^{k} (-1)^{k-j} \sum_{(t_1,\dots,t_k)\in\{0,1\}^k, \sum_i t_i = j} \mathbb{E} f(G_k(\theta; t_1, \dots, t_k))$$

$$= \mathbb{E} \sum_{(t_1,\dots,t_k)\in\{0,1\}^k} (-1)^{k-\sum_{i=1}^k t_i} f(G_k(\theta; t_1, \dots, t_k))$$

$$= \mathbb{E}\Delta^{(1)} \dots \Delta^{(k)} f(G_k(\theta; t_1, \dots, t_k))$$

$$= \mathbb{E} \int_0^1 \dots \int_0^1 \frac{\partial^k f(G_k(\theta; t_1, \dots, t_k))}{\partial t_1 \dots \partial t_k} dt_1 \dots dt_k$$

$$= \mathbb{E} \frac{\partial^k f(G_k(\theta; U_1, \dots, U_k))}{\partial t_1 \dots \partial t_k}, \ U_1, \dots, U_k \text{ i.i.d. } U[0,1].$$

Francesco Faà di Bruno

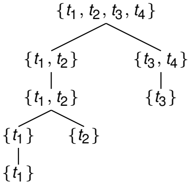
Faà di Bruno's Formula (1855):

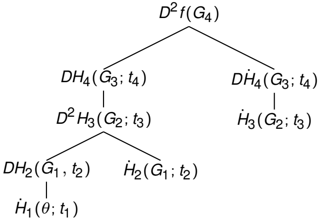
$$\frac{d^n}{dx^n} f(g(x)) = \sum \frac{n!}{k_1! \dots k_n!} f^{(k_1+\cdots+k_n)}(g(x)) \prod_{j=1}^{n} \left(\frac{g^{(j)}(x)}{j!} \right)^{k_j},$$

where the sum is over all $(k_1, \dots, k_n)$, $k_j \geq 0$, $\sum_{j=1}^{n} j k_j = n$.

A representation formula for $\mathcal{B}^k f$

$$(\mathcal{B}^k f)(\theta) = \sum_{\tau \in \mathcal{T}_k} \mathbb{E}\partial_\tau f(G_k(\theta; U_1, \dots, U_k)), \theta \in \Theta.$$





For this tree, $\partial_\tau f(G_4) = D^2 f(G_4)[DH_4(G_3; t_4)[D^2 H_3(G_2; t_3)[DH_2(G_1; t_2)[\dot{H}_1(\theta; t_1)], \dot{H}_2(G_1; t_2)]], D\dot{H}_4(G_3; t_4)[\dot{H}_3(G_2; t_3)]]$.

Random Homotopies and Representation of Operators $\mathcal{B}^k$

If functional f and random homotopy H are sufficiently smooth, “Faà di Bruno type calculus" could be used to obtain representations for $\mathcal{B}^k f(\theta)$ and to prove upper bounds

$$|B^k f(\theta)| \lesssim$$

$$\|f\|_{C^k}\biggl(\mathbb{E}\Bigl(\sup_{t\in[0,1]}\|H(\cdot;t)\|_{C^{k-1}}\vee 1\Bigr)^k \sup_{t\in[0,1]}\Bigl\|\frac{d}{dt}H(\cdot;t)\Bigr\|_{C^{k-1}}\biggr)^k$$

of the order $O\Bigl(\Bigl(\sqrt{\frac{d}{n}}\Bigr)^k\Bigr)$ provided that

$$\sup_{t\in[0,1]}\Bigl\|\frac{d}{dt}H(\cdot;t)\Bigr\|_{C^{k-1}} \lesssim \sqrt{\frac{d}{n}}.$$

One can also study smoothness properties of functional f_k and derive concentration and normal approximation bounds for $f_k(\hat{\theta})$.

Beyond the Gaussian Case, K (2020)

- $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in E$, E a Banach space
- $\hat\theta(X_1, \dots, X_n)$ an estimator of θ
- $\sqrt{n}(\hat\theta - \theta) \overset{d}{\approx} \xi(\theta)$
- $\xi : E \mapsto E$ a Gaussian process
- $\mathbb{E}\|\xi\|_{C^s}^2 \lesssim d$
- If $E = \mathbb{R}^d$, $\hat\theta$ is MLE, then $\xi(\theta) = A(\theta)Z$, $Z \sim N(0; I_d)$, where $A(\theta) = I(\theta)^{-1/2}$, $I(\theta)$ the Fisher information
- If $\|A\|_{C^s} \lesssim 1$, then $\mathbb{E}\|\xi\|_{C^s}^2 \lesssim d$

Gaussian Approximation

Assumption

(NA_s). Let $s > 1$. Suppose that

$$\sup_{\theta \in E} \sup_{\|g'\|_{C^{s-1}} \leq 1} |\mathbb{E}_\theta g(\sqrt{n}(\hat{\theta} - \theta)) - \mathbb{E} g(\xi(\theta))| \to 0$$

as $n \to \infty$.

Let

$$\sigma_f^2(\theta) := \mathbb{E}\langle \xi(\theta), f'(\theta)\rangle^2.$$

Asymptotic Normality

Theorem

Suppose $d = d_n \leq n^\alpha$ *for some* $\alpha \in (0,1)$. *Let* $s = k + 1 + \rho > \frac{1}{1-\alpha}$, $k \geq 0$, $\rho \in (0,1]$ *and suppose that condition* (NA_s) *holds. Then*

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\theta \in E} \left| n\mathbb{E}_\theta (f_k(\hat{\theta}) - f(\theta))^2 - \sigma_f^2(\theta) \right| \to 0, \ \ n \to \infty$$

and, for all $\sigma_0 > 0$ *and for* $Z \sim N(0,1)$,

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\theta \in E, \sigma_f(\theta) \geq \sigma_0} d_K\left(\frac{\sqrt{n}(f_k(\hat{\theta}) - f(\theta))}{\sigma_f(\theta)}; Z \right) \to 0, n \to \infty.$$

Example

- $X = \theta + \eta \in \mathbb{R}^d$, $\eta \sim \mu_\theta$, $\eta \overset{d}{=} -\eta$
- $X_1, \dots, X_n$ i.i.d. copies of X
- $\hat{\theta} := \bar{X} = \frac{X_1 + \cdots + X_n}{n}$
- μ a probability measure on $\mathbb{R}^d$
- Let $C_P(\mu)$ be the Poincaré constant of μ :

$$C_P(\mu) := \inf\Big\{C > 0 : \mathrm{Var}_\mu(g(X)) \leq C\mathbb{E}_\mu \|(\nabla g)(X)\|^2 \\ \forall g : \mathbb{R}^d \mapsto \mathbb{R} \text{ locally Lipschitz}\Big\}$$

- Suppose $\sup_{\theta \in \mathbb{R}^d} C_P(\mu_\theta) < \infty$
- $\Sigma(\theta) = \mathbb{E}_\theta (X - \theta) \otimes (X - \theta)$
- $\|\Sigma\|_{C^s} \lesssim 1$, $\|\Sigma^{-1}\|_{L_\infty} \lesssim 1$

Example

Corollary

Suppose $d = d_n \leq n^{\alpha}$ *for some* $\alpha \in (0,1)$. *Let* $s = k + 1 + \rho > \frac{1}{1-\alpha}$, $k \geq 0$, $\rho \in (0,1]$ *and*

$$\sup_{\theta \in \mathbb{R}^d} C_P(\mu_\theta) = o(n^{1-\alpha}) \text{ as } n \to \infty.$$

Then

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\theta \in E} \left| n\mathbb{E}_\theta (f_k(\hat{\theta}) - f(\theta))^2 - \sigma_f^2(\theta) \right| \to 0, \ n \to \infty$$

and, for all $\sigma_0 > 0$ *and for* $Z \sim N(0,1)$,

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\theta \in E, \sigma_f(\theta) \geq \sigma_0} d_K\left(\frac{\sqrt{n}(f_k(\hat{\theta}) - f(\theta))}{\sigma_f(\theta)}; Z \right) \to 0, n \to \infty.$$

Example

- The proof is based on recent bounds on convergence rates in high-dimensional CLT by Courtade, Fathi and Pananjadi (2018), Fathi (2018)
- If μ_θ is log-concave, then the condition on Poincarè constant holds if $\alpha < 2/3$ (by the bounds by Lee and Vempala (2017)) and it, most likely, holds for all $\alpha \in (0, 1)$ (subject to KLS conjecture).