Weierstrass Institute for
Applied Analysis and Stochastics

Leibniz
Association

Random gradient free optimization: Bayesian view

Vladimir Spokoiny ,
WIAS, HU Berlin

April, 2020

Aim: An efficient procedure to minimize a convex function $f(\boldsymbol{x})$ without computing the gradient and Hessian.

[Nesterov and Spokoiny, 2017] offered a "gradient free" procedure which only relies on the directional derivative of $f(\boldsymbol{x})$

Procedure: with $\boldsymbol{u} \sim \mathcal{N}(0, B^{-1})$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k \frac{f(\boldsymbol{x}_k + \mu \boldsymbol{u}) - f(\boldsymbol{x})}{\mu} \cdot B\boldsymbol{u}.$$

The basic tool of study is the average

$$f_\mu(\boldsymbol{x}) = \mathbb{E} f(\boldsymbol{x} + \mu \boldsymbol{u}) = \frac{\int f(\boldsymbol{x} + \mu \boldsymbol{u}) \mathrm{e}^{-\langle B\boldsymbol{u}, \boldsymbol{u}\rangle/2} \, d\boldsymbol{u}}{\int \mathrm{e}^{-\langle B\boldsymbol{u}, \boldsymbol{u}\rangle/2} \, d\boldsymbol{u}}$$

Let $\boldsymbol{Y}$ denote the observed random data, $\boldsymbol{Y} \sim I\!\!P$.

Model (DNN): $I\!\!P \in (I\!\!P_{\boldsymbol{\upsilon}}, \boldsymbol{\upsilon} \in \varUpsilon \subseteq I\!\!R^{\infty})$.

The log-likelihood function (negative fidelity)

$$L(\boldsymbol{\upsilon}) = L(\boldsymbol{Y}, \boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \log \frac{dI\!\!P_{\boldsymbol{\upsilon}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y}).$$

Training by MLE: maximizing the random function $L(\boldsymbol{\upsilon})$

$$\widetilde{\boldsymbol{\upsilon}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} L(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} \exp L(\boldsymbol{\upsilon}).$$

Target (the best parametric fit/ risk minimization):

$$\boldsymbol{\upsilon}^{*} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} I\!\!E L(\boldsymbol{\upsilon}) \neq \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} I\!\!E \exp L(\boldsymbol{\upsilon}).$$

Also concavity of $L(\boldsymbol{\upsilon})$ $\neq$ concavity of $\exp L(\boldsymbol{\upsilon})$

- Υ is high or infinite dimensional
- $\nabla L(\boldsymbol{\upsilon})$ and $\nabla^2 L(\boldsymbol{\upsilon})$ hard to compute
- badly posed, need of regularization
- non-convex and non-smooth problem
- dimension reduction issue (drop-out)
- standard technique: SGD + backpropagation; efficient numerically but theoretical guarantees are hard to obtain

Bayesian (MCMC-type) methods with Gaussian priors yielding

- Efficient gradient and Hessian free procedure with second order accuracy
- Theoretical guarantees

Outline

In the Bayes setup $\boldsymbol{\upsilon}$ is a random element, $\boldsymbol{\upsilon} \sim \varPi$ on the parameter set $\varUpsilon$, a prior with density $\varPi(\boldsymbol{\upsilon})$.

The posterior describes the conditional distribution of $\boldsymbol{\upsilon}$ given $\boldsymbol{Y}$

$$\boldsymbol{\upsilon} \, \big| \, \boldsymbol{Y} \sim \frac{\exp\bigl\{ L(\boldsymbol{\upsilon}) \bigr\} \, \varPi(\boldsymbol{\upsilon})}{\int \exp\bigl\{ L(\boldsymbol{\upsilon}) \bigr\} \, \varPi(\boldsymbol{\upsilon}) d\boldsymbol{\upsilon}} = \frac{\exp\bigl\{ L_{\varPi}(\boldsymbol{\upsilon}) \bigr\}}{\int \exp\bigl\{ L_{\varPi}(\boldsymbol{\upsilon}) \bigr\} \, d\boldsymbol{\upsilon}}$$

with $L_{\varPi} = L(\boldsymbol{\upsilon}) + \log \varPi(\boldsymbol{\upsilon})$.

Prior $\varPi$ induces a penalty $-\log \varPi(\boldsymbol{\upsilon})$ leading to penalized MLE

$$\widetilde{\boldsymbol{\upsilon}}_{\varPi} = \operatorname*{argmax}_{\boldsymbol{\upsilon}} L_{\varPi}(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\upsilon}} \bigl\{ L(\boldsymbol{\upsilon}) + \log \varPi(\boldsymbol{\upsilon}) \bigr\}.$$

Formally, Bayes approach replaces the max of $L_{\varPi}$ by the soft-max.

A Gaussian priors $\mathcal{N}(\bar{\boldsymbol{v}}, G^{-2})$ lead to quadratic penalization

$$\widetilde{\boldsymbol{v}}_G = \operatorname*{argmax}_{\boldsymbol{v}} L_G(\boldsymbol{v}) = \operatorname*{argmax}_{\boldsymbol{v}} \bigl\{ L(\boldsymbol{v}) - \bigl\| G(\boldsymbol{v} - \bar{\boldsymbol{v}}) \bigr\|^2 / 2 \bigr\};$$

cf. the Moreau–Yosida proximal-point method.

Posterior $\boldsymbol{v}_G \, \big| \, \boldsymbol{Y}$ is a random measure with the density

$$\boldsymbol{v}_G \, \big| \, \boldsymbol{Y} \sim \frac{\exp L_G(\boldsymbol{v})}{\int \exp L_G(\boldsymbol{v}) \, d\boldsymbol{v}} \propto \exp L_G(\boldsymbol{v}).$$

1. Select a starting prior $\mathit{\Pi}_0 = \mathcal{N}(\bar{\boldsymbol{v}}_0, G_0^{-2})$; Set $k = 0$;
2. Draw an independent sample $\boldsymbol{v}_1^{(k)}, \ldots, \boldsymbol{v}_M^{(k)}$ from the prior $\mathcal{N}\bigl(\bar{\boldsymbol{v}}_k, G_k^{-2}\bigr)$. For each $\boldsymbol{v}_m^{(k)}$, compute $L_k(\boldsymbol{v}_m^{(k)})$ with

$$L_k(\boldsymbol{v}) = L(\boldsymbol{v}) - \|G_k(\boldsymbol{v} - \bar{\boldsymbol{v}}_k)\|^2/2$$

   and the corresponding weight

$$w_m^{(k)} = \exp L_k(\boldsymbol{v}_m^{(k)}).$$

3. Use the collection $(\boldsymbol{v}_m^{(k)}, w_m^{(k)})$ for $m \leq M$ (posterior) to build the next prior distribution $\mathit{\Pi}_{k+1} = \mathcal{N}(\bar{\boldsymbol{v}}_{k+1}, G_{k+1}^{-2})$.
4. Increase $k \to k + 1$ and repeat pp. 2 and 3 until convergence.

A prior is called conjugated if the corresponding posterior belongs to the same family of measures as prior. This reduces the step of computing the posterior to parameter update.

Our main result claims that the Gaussian prior for a regular parametric family yields a nearly Gaussian posterior (nearly conjugated). Therefore, it suffices to recompute the parameters of normal law.

It is natural and standard to use the posterior mean

$$\overline{\boldsymbol{\upsilon}}_{k+1} = \widehat{\boldsymbol{\upsilon}}_k = \frac{1}{N_k} \sum_m w_m^{(k)} \boldsymbol{\upsilon}_m^{(k)} , \qquad N_k = \sum_m w_m^{(k)} .$$

Alternatively, a robust (trimmed) mean could be used.

We suggest to apply $G_k^2 = \rho_k^{-2} G^2$ with a fixed G^2.

The prior concentrates on the ρ_k-vicinity of $\overline{\boldsymbol{\upsilon}}_k$ and the same for posterior, so, ρ_k has the flavor of a step size.

Let $\mathbb{E}L_k(\boldsymbol{\upsilon})$ be (locally) concave. Define $D_k^2 = -\nabla^2 \mathbb{E}L_k(\overline{\boldsymbol{\upsilon}}_k)$ and consider a quadratic approximation of $L_k(\boldsymbol{\upsilon})$ around $\overline{\boldsymbol{\upsilon}}_k$.

As $\widetilde{\boldsymbol{\upsilon}}_k = \operatorname{argmax}_{\boldsymbol{\upsilon}} L_k(\boldsymbol{\upsilon})$, it holds

$$
\begin{gathered}
L_k(\boldsymbol{\upsilon}) - L_k(\widetilde{\boldsymbol{\upsilon}}_k) \approx -\bigl\| D_k(\boldsymbol{\upsilon} - \widetilde{\boldsymbol{\upsilon}}_k) \bigr\|^2 / 2, \\
\nabla L_k(\widetilde{\boldsymbol{\upsilon}}_k) = 0, \\
\nabla L_k(\overline{\boldsymbol{\upsilon}}_k) - \nabla L_k(\widetilde{\boldsymbol{\upsilon}}_k) \approx -D_k^2(\overline{\boldsymbol{\upsilon}}_k - \widetilde{\boldsymbol{\upsilon}}_k),
\end{gathered}
$$

and hence $\overline{\boldsymbol{\upsilon}}_{k+1} = \widetilde{\boldsymbol{\upsilon}}_k$ corresponds to the Newton update:

$$
\overline{\boldsymbol{\upsilon}}_{k+1} - \overline{\boldsymbol{\upsilon}}_k \approx D_k^{-2} \nabla L_k(\overline{\boldsymbol{\upsilon}}_k).
$$

The use of $\nabla L_k(\widetilde{\boldsymbol{\upsilon}}_k) = 0$ yields with $\widetilde{D}_k^2 = -\nabla^2 \mathbb{E} L_k(\widetilde{\boldsymbol{\upsilon}}_k)$

$$\frac{\int g(\boldsymbol{u}) \exp\bigl\{ L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) \bigr\} \, d\boldsymbol{u}}{\int \exp\bigl\{ L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) \bigr\} \, d\boldsymbol{u}}$$

$$= \frac{\int g(\boldsymbol{u}) \exp\bigl\{ L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G) \bigr\} \, d\boldsymbol{u}}{\int \exp\bigl\{ L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G) \bigr\} \, d\boldsymbol{u}}$$

$$= \frac{\int g(\boldsymbol{u}) \, \exp\bigl\{ L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{\upsilon}}_G), \boldsymbol{u} \bigr\rangle \bigr\} d\boldsymbol{u}}{\int \exp\bigl\{ L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{\upsilon}}_G), \boldsymbol{u} \bigr\rangle \bigr\} d\boldsymbol{u}}$$

$$\approx \frac{\int g(\boldsymbol{u}) \, \exp\bigl\{ -\| \widetilde{D}_k \boldsymbol{u} \|^2 / 2 \bigr\} \, d\boldsymbol{u}}{\int \exp\bigl\{ -\| \widetilde{D}_k \boldsymbol{u} \|^2 / 2 \bigr\} \, d\boldsymbol{u}}$$

hence $\boldsymbol{\upsilon}_k \bigm| \boldsymbol{Y} \rightsquigarrow \mathcal{N}(\widetilde{\boldsymbol{\upsilon}}_k, \widetilde{D}_k^{-2})$.

The procedure is gradient and Hessian free. Each step delivers a posterior distribution $\boldsymbol{\upsilon}_k \,\big|\, \boldsymbol{Y}$. We show that

$$\boldsymbol{\upsilon}_k \,\big|\, \boldsymbol{Y} \approx \mathcal{N}(\widetilde{\boldsymbol{\upsilon}}_k, \widetilde{D}_k^{-2}), \quad \widetilde{D}_k^2 = D_k^2(\widetilde{\boldsymbol{\upsilon}}_k)$$

and, in particular,

- $\boldsymbol{\upsilon}_k \,\big|\, \boldsymbol{Y}$ concentrates on an elliptic vicinity of $\widetilde{\boldsymbol{\upsilon}}_k$
- posterior mean $\widehat{\boldsymbol{\upsilon}}_k$ is a proxi for $\widetilde{\boldsymbol{\upsilon}}_k$ and can be computed from Bayesian sampling: $\widehat{\boldsymbol{\upsilon}}_k = N_k^{-1} \sum_m \boldsymbol{\upsilon}_m^{(k)} w_m^{(k)}$;
- Information from all previous steps can be incorporated in the Gaussian prior $\mathcal{N}(\overline{\boldsymbol{\upsilon}}_k, G_k^{-2})$;
- Prior precision matrix G_k^2 can be used to manipulate with the step size and effective dimension.

Even in the parametric case, the value $\boldsymbol{\upsilon}^{*}$ can be estimated with accuracy $n^{-1/2}$ at best.

Therefore, no sense to continue the procedure if $D_k^{-2} \ll n^{-1}$.

If ρ_k decreases exponentially, only $\log n$ steps are required.

$(\boldsymbol{E})$ *The stochastic component* $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E}L(\boldsymbol{\upsilon})$ *of the process* $L(\boldsymbol{\upsilon})$ *is linear in* $\boldsymbol{\upsilon}$ *:*

$$\nabla\zeta \equiv \nabla\zeta(\boldsymbol{\upsilon}).$$

$(\boldsymbol{ED_0})$ *There exist a positive symmetric matrix* V *, and constants* $\mathrm{g} > 0$ *,* $\nu_0 \geq 1$ *such that* $\mathrm{Var}\bigl(\nabla\zeta\bigr) \leq V^2$ *and*

$$\sup_{\boldsymbol{u}\in\mathbb{R}^p} \log \mathbb{E} \exp\biggl\{ \lambda \frac{\langle \boldsymbol{u}, \nabla\zeta \rangle}{\| V\boldsymbol{u} \|} \biggr\} \leq \frac{\nu_0^2 \lambda^2}{2}, \qquad |\lambda| \leq \mathrm{g}.$$

$(\mathcal{L})$ *The set Υ is open and convex in $\mathbb{R}^p$. For each k, the function $\mathbb{E}L_k(\boldsymbol{\upsilon})$ is concave in $\boldsymbol{\upsilon} \in \Upsilon$.*

$(\mathcal{L}_0)$ Define for each $\boldsymbol{\upsilon} \in \Upsilon^\circ$, and any $\boldsymbol{u} \in \mathbb{R}^p$, the directional derivative

$$\delta_3(\boldsymbol{\upsilon}, \boldsymbol{u}) \stackrel{\text{def}}{=} \frac{1}{m!} \frac{d^m}{dt^m} \mathbb{E}L(\boldsymbol{\upsilon} + t\boldsymbol{u}) \Big|_{t=0}, \quad m = 3, 4.$$

The functions $\delta_3(\boldsymbol{\upsilon}, \boldsymbol{u})$ and $\delta_4(\boldsymbol{\upsilon}, \boldsymbol{u})$ are well defined and with $D^2(\boldsymbol{\upsilon}) = -\nabla^2 \mathbb{E}L(\boldsymbol{\upsilon})$

$$\omega_m(\boldsymbol{\upsilon}) = \sup_{\boldsymbol{u} : \, \|D(\boldsymbol{\upsilon})\boldsymbol{u}\| \le \mathtt{r}} \frac{\delta_m(\boldsymbol{\upsilon}, \boldsymbol{u})}{\|D(\boldsymbol{\upsilon})\boldsymbol{u}\|^2} \le \frac{1}{3}.$$

Define

$$\mathbb{F}(\boldsymbol{\upsilon}) = -\nabla^2 \mathbb{E} L(\boldsymbol{\upsilon}), \quad \mathbb{F}_G(\boldsymbol{\upsilon}) = \mathbb{F}(\boldsymbol{\upsilon}) + G^2.$$

$(\boldsymbol{V}|\boldsymbol{G})$ Signal-noise:

$$B_{V|G}(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \mathbb{F}_G^{-1/2}(\boldsymbol{\upsilon})\, V^2\, \mathbb{F}_G^{-1/2}(\boldsymbol{\upsilon})$$

with V^2 from $(\boldsymbol{ED_0})$. There are fixed constants $\lambda_{V|G}$ and $\mathrm{p}_{V|G}$ such that

$$\operatorname{tr} B_{V|G}(\boldsymbol{\upsilon}) \le \mathrm{p}_{V|G}\,, \qquad \|B_{V|G}\| \le \lambda_{V|G}\,, \qquad \boldsymbol{\upsilon} \in \Upsilon^{\circ}\,.$$

$(\boldsymbol{D}|\boldsymbol{G})$ The effective dimension: for a fixed constant $\mathtt{C}$

$$\mathrm{p}_G(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \operatorname{tr}\bigl\{\mathbb{F}(\boldsymbol{\upsilon})\mathbb{F}_G^{-1}(\boldsymbol{\upsilon})\bigr\} \le \mathtt{C}, \qquad \boldsymbol{\upsilon} \in \Upsilon^{\circ}.$$

Consider the model

$$\boldsymbol{Y} = A(\boldsymbol{f}) + \sigma\boldsymbol{\varepsilon} \in \mathcal{Y}^d$$

with a known non-linear operator $A\colon \mathcal{X} \to \mathcal{Y}^d$ for Hilbert spaces $\mathcal{X}, \mathcal{Y}$ and a discretized subspace $\mathcal{Y}^d \subset \mathcal{Y}$.

Examples include

- PDE with elliptic operators, [Nickl et al., 2018]
- Schrödinger equation, [Nickl, 2017]
- Calderón equation, [Abraham and Nickl, 2019]

$$\boldsymbol{f} \,\big|\, \boldsymbol{Y} \propto \exp L_G(\boldsymbol{f}),$$

$$L_G(\boldsymbol{f}) = -\frac{1}{2\sigma^2}\|\boldsymbol{Y} - A(\boldsymbol{f})\|^2 - \frac{1}{2}\|G\boldsymbol{f}\|^2 .$$

Let observations Y_i, X_i follow the model

$$Y_i \sim P_{f(X_i)} \in \bigl(P_u\bigr),$$

where (P_u) is an exponential family with a log-density

$$\ell(y, u) = \log p(y, u) = C(u)y - B(u);$$

$C(u)$ is an increasing and $B(u)$ is a convex function.

Include binary-response, Poissonian regression, Cox regression, reliability and extreme values, ...

PA (DNN): $f(x) = f(x, \boldsymbol{\upsilon}^*)$. Yields the log-likelihood

$$L(\boldsymbol{\upsilon}) = \sum_{i=1}^{n} \ell(Y_i, f(X_i, \boldsymbol{\upsilon})) = \sum_{i=1}^{n} Y_i \bigl\{ C\bigl(f(X_i, \boldsymbol{\upsilon})\bigr) - B\bigl(f(X_i, \boldsymbol{\upsilon})\bigr) \bigr\}.$$

- MAP and posterior mean

$$\widetilde{\boldsymbol{v}}_{\mathrm{MAP}} = \operatorname*{argmax}_{\boldsymbol{v}} \exp L_G(\boldsymbol{v}) = \widetilde{\boldsymbol{v}}_G$$

$$\overline{\boldsymbol{v}}_G \stackrel{\mathrm{def}}{=} \frac{\int \boldsymbol{v} \exp L_G(\boldsymbol{v}) d\boldsymbol{v}}{\int \exp L_G(\boldsymbol{v}) d\boldsymbol{v}} \, ;$$

- Concentration set $\mathcal{A}$:

$$\rho(\mathcal{A}) \stackrel{\mathrm{def}}{=} \frac{\int_{\mathcal{A}^c} \exp L_G(\boldsymbol{v}) d\boldsymbol{v}}{\int_{\mathcal{A}} \exp L_G(\boldsymbol{v}) d\boldsymbol{v}}$$

- Credible sets $\mathcal{A}(\alpha)$

$$\mathbb{P}\bigl(\boldsymbol{v}_G \in \mathcal{A}(\alpha) \bigm| \boldsymbol{Y}\bigr) = 1 - \alpha;$$

- Elliptic credible sets $\mathcal{A}_{Q|G}(\alpha) = \bigl\{\boldsymbol{v} \colon \bigl\| Q\bigl(\boldsymbol{v}_G - \widetilde{\boldsymbol{v}}_G\bigr)\bigr\| \le z_\alpha \bigr\}$.

Penalized MLE (pMLE): with $L_G(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \|G\boldsymbol{\upsilon}\|^2/2$

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmax}_{\boldsymbol{\upsilon}} L_G(\boldsymbol{\upsilon}), \qquad \boldsymbol{\upsilon}_G^* = \operatorname*{argmax}_{\boldsymbol{\upsilon}} \mathbb{E} L_G(\boldsymbol{\upsilon})$$

Full information matrix (operator)

$$\mathbb{F}(\boldsymbol{\upsilon}) = -\nabla^2 \mathbb{E} L(\boldsymbol{\upsilon}),$$

$$\mathbb{F}_G(\boldsymbol{\upsilon}) = -\nabla^2 \mathbb{E} L_G(\boldsymbol{\upsilon}) = \mathbb{F}(\boldsymbol{\upsilon}) + G^2 \, ,$$

and

$$D_G^2 = \mathbb{F}_G(\boldsymbol{\upsilon}_G^*), \qquad \widetilde{D}_G^2 = \mathbb{F}_G(\widetilde{\boldsymbol{\upsilon}}_G).$$

Effective dimension

$$\mathrm{p}_G(\boldsymbol{\upsilon}) = \operatorname{tr}\Bigl(\mathbb{F}_G^{-1}(\boldsymbol{\upsilon}) \mathbb{F}(\boldsymbol{\upsilon})\Bigr), \quad \mathrm{p}_G = \mathrm{p}_G(\boldsymbol{\upsilon}_G^*), \quad \widetilde{\mathrm{p}}_G = \mathrm{p}_G(\widetilde{\boldsymbol{\upsilon}}_G).$$

Fisher expansion for pMLE: on a set Ω of high probability

$$\left\| D_G\left(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*_G\right) - D_G^{-1}\nabla\zeta\right\|^2 \lesssim \sqrt{\frac{\mathtt{p}_G}{n}}\, \|D_G^{-1}\nabla\zeta\|^2$$

with $\nabla\zeta = \nabla L_G(\boldsymbol{\upsilon}^*_G)$.

Wilks expansion

$$\left| 2L_G(\widetilde{\boldsymbol{\upsilon}}_G) - 2L_G(\boldsymbol{\upsilon}^*_G) - \left\| D_G^{-1}\nabla\zeta\right\|^2 \right| \lesssim \sqrt{\frac{\mathtt{p}_G}{n}}\, \left\| D_G^{-1}\nabla\zeta\right\|^2,$$

$$\left| 2L_G(\widetilde{\boldsymbol{\upsilon}}_G) - 2L_G(\boldsymbol{\upsilon}) - \left\| D_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon})\right\|^2 \right| \lesssim \sqrt{\frac{\mathtt{p}_G}{n}}\, \left\| D_G^{-1}\nabla\zeta\right\|^2.$$

$$\sup_{A \in \mathcal{B}_s(\mathbb{R}^p)} \left| \mathbb{P}\big(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \in A \,\big|\, \boldsymbol{Y}\big) - \mathbb{P}'\big(\widetilde{D}_G^{-1} \boldsymbol{\gamma} \in A\big) \right| \lesssim \frac{\mathrm{p}_G^3}{n}$$

$$\sup_{A \in \mathcal{B}(\mathbb{R}^p)} \left| \mathbb{P}\big(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \in A \,\big|\, \boldsymbol{Y}\big) - \mathbb{P}'\big(\widetilde{D}_G^{-1} \boldsymbol{\gamma} \in A\big) \right| \lesssim \sqrt{\frac{\mathrm{p}_G^3}{n}}$$

where $\mathcal{B}(\mathbb{R}^p)$ stands for all Borel sets while $\mathcal{B}_s(\mathbb{R}^p)$ all centrally symmetric Borel sets in $\mathbb{R}^p$.

- rate of estimation of pMLE
- posterior concentration and contraction rate
- use of posterior mean in place of MAP
- credible sets as frequentist confidence sets
- prior impact
- empirical or full Bayes approach for prior selection

All for

- finite samples
- explicit error terms via effective dimension instead of full parameter dimension

Most of results requires the upper bound on the effective dimension $\mathrm{p}_G = \operatorname{tr}\left(\mathbb{F}_G^{-1}\mathbb{F}\right)$

$$\mathrm{p}_G \ll n$$

However, the main result on Gaussian approximation of the posterior only valid under

$$\mathrm{p}_G^3 \ll n$$

By $(\boldsymbol{E})$, the stochastic component $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - I\!\!E L(\boldsymbol{\upsilon})$ is linear in $\boldsymbol{\upsilon}$ and $\nabla\zeta = \nabla\zeta(\boldsymbol{\upsilon})$ does not depend on $\boldsymbol{\upsilon}$.

Theorem Under condition $(\boldsymbol{ED_0})$, there exists a random set $\Omega(\mathtt{x})$ with $I\!\!P\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$ such that on this set

$$
\bigl\| D_G^{-1} \nabla\zeta \bigr\| \leq z(B_{V|G}, \mathtt{x}),
$$

where $B_{V|G} = D_G^{-1} V^2 D_G^{-1}$ and

$$
z(B_{V|G}, \mathtt{x}) = \sqrt{\operatorname{tr} B_{V|G}} + \sqrt{2\mathtt{x} \lambda_{\max}(B_{V|G})}.
$$

Let $\boldsymbol{\upsilon}_G^* = \operatorname{arginf}_{\boldsymbol{\upsilon}} \mathbb{E} L_G(\boldsymbol{\upsilon})$ and $D_G^2 = D^2(\boldsymbol{\upsilon}_G^*) + G^2$.

Theorem Let $\left\| D_G^{-1} \nabla \zeta \right\| \le z(B_{V|G}, \mathtt{x})$ on a random set $\Omega(\mathtt{x})$ with $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \ge 1 - \mathrm{e}^{-\mathtt{x}}$. Let

$$
\begin{aligned}
\mathcal{A}_G(\mathtt{r}_G) &\stackrel{\text{def}}{=} \bigl\{ \boldsymbol{\upsilon} \colon \| D_G(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_G^*) \| \le \mathtt{r}_G \bigr\} \\
(1 - \rho)\mathtt{r}_G &\ge z(B_{V|G}, \mathtt{x}).
\end{aligned}
$$

Then on $\Omega(\mathtt{x})$

$$
\left\| D_G \bigl(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^* \bigr) \right\| \le \mathtt{r}_G .
$$

Local set

$$\mathcal{A}_G(\mathtt{r}_G) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\upsilon}\colon \|D_G(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_G^*)\| \le \mathtt{r}_G\bigr\}.$$

Use local smoothness of $-\mathbb{E}L_G(\boldsymbol{\upsilon})$ to show that for $\boldsymbol{\upsilon} \in \mathcal{A}_G(\mathtt{r}_G)$

$$-\bigl\{\mathbb{E}L_G(\boldsymbol{\upsilon}) - \mathbb{E}L_G(\boldsymbol{\upsilon}_G^*)\bigr\} \approx \bigl\|D_G(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_G^*)\bigr\|^2/2,$$

$$-\nabla\mathbb{E}L_G(\boldsymbol{\upsilon}) \approx D_G^2(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_G^*)$$

Use convexity of $\mathbb{E}L_G(\boldsymbol{\upsilon})$ to show that

$$\bigl\|\nabla\mathbb{E}L_G(\boldsymbol{\upsilon})\bigr\| \ge \mathtt{r}_G\,, \quad \boldsymbol{\upsilon} \notin \mathcal{A}_G(\mathtt{r}_G).$$

Use $\bigl\|D_G^{-1}\nabla\zeta\bigr\| \le z(B_{V|G}, \mathtt{x})$ to show

$$\nabla L(\boldsymbol{\upsilon}) = \nabla\mathbb{E}L(\boldsymbol{\upsilon}) + \nabla\zeta \ne 0, \quad \boldsymbol{\upsilon} \notin \mathcal{A}_G(\mathtt{r}_G).$$

Concentration sets of the posterior in nonparametric models using

- Empirical process theory, large deviations of the log-likelihood and covering numbers and chaining arguments
- small ball probability
- local smoothness

See e.g.

- Ghoshal, S., Ghosh, J. K., van der Vaart, A. W. (2000) Convergence rates of posterior distributions. Ann.Statist., 28, 500–531.
- van der Vaart, A. W., van Zanten, J. H. (2008) Rates of contraction of posterior distributions based on Gaussian process priors. Ann. Statist., 36, 1031–1508.

Define the elliptic set $\bigl\{ \boldsymbol{u} \colon \|\widetilde{D}\boldsymbol{u}\| \le \mathtt{r}_0 \bigr\}$ with $\widetilde{D}^2 = D^2(\widetilde{\boldsymbol{\upsilon}}_G)$.
Consider the random quantity

$$\rho(\mathtt{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\|\widetilde{D}\boldsymbol{u}\| > \mathtt{r}_0} \exp\bigl\{ L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) \bigr\} d\boldsymbol{u}}{\int_{\|\widetilde{D}\boldsymbol{u}\| \le \mathtt{r}_0} \exp\bigl\{ L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) \bigr\} d\boldsymbol{u}} .$$

Theorem Let, for some fixed values $\mathtt{r}_0$ and $\mathtt{x} > 0$, it hold

$$\mathtt{C}_0 \mathtt{r}_0 \ge 2\sqrt{\mathrm{p}_G(\boldsymbol{\upsilon})} + \sqrt{\mathtt{x}}\,, \qquad \boldsymbol{\upsilon} \in \Upsilon^{\circ} .$$

Then, on the random set $\Omega(\mathtt{x})$ from Theorem 31, with $\widetilde{\mathrm{p}}_G = \mathrm{p}_G(\widetilde{\boldsymbol{\upsilon}}_G)$

$$\rho(\mathtt{r}_0) \le \exp\bigl\{ -(\widetilde{\mathrm{p}}_G + \mathtt{x})/2 \bigr\}.$$

Let $\widetilde{\boldsymbol{v}}_G = \operatorname{arginf}_{\boldsymbol{v}} L_G(\boldsymbol{v})$. The use of $\nabla L_G(\widetilde{\boldsymbol{v}}_G) = 0$ allows to represent

$$
\begin{aligned}
\rho(\mathtt{r}_0) &= \frac{\int_{\|\widetilde{D}\boldsymbol{u}\| > \mathtt{r}_0} \exp\bigl\{L_G(\widetilde{\boldsymbol{v}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{v}}_G)\bigr\} d\boldsymbol{u}}{\int_{\|\widetilde{D}\boldsymbol{u}\| \le \mathtt{r}_0} \exp\bigl\{L_G(\widetilde{\boldsymbol{v}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{v}}_G)\bigr\} d\boldsymbol{u}} \\
&= \frac{\int_{\|\widetilde{D}\boldsymbol{u}\| > \mathtt{r}_0} \exp\bigl\{L_G(\widetilde{\boldsymbol{v}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{v}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{v}}_G), \boldsymbol{u} \bigr\rangle\bigr\} d\boldsymbol{u}}{\int_{\|\widetilde{D}\boldsymbol{u}\| \le \mathtt{r}_0} \exp\bigl\{L_G(\widetilde{\boldsymbol{v}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{v}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{v}}_G), \boldsymbol{u} \bigr\rangle\bigr\} d\boldsymbol{u}}
\end{aligned}
$$

Fix $\boldsymbol{v} \in \mathcal{A}_G(\mathtt{r}_G)$. Consider $f(\boldsymbol{v}) = \mathbb{E} L_G(\boldsymbol{v})$. As $\zeta(\boldsymbol{v}) = L(\boldsymbol{v}) - \mathbb{E} L(\boldsymbol{v})$ is linear in $\boldsymbol{v}$, it holds

$$L_G(\boldsymbol{v}+\boldsymbol{u}) - L_G(\boldsymbol{v}) - \bigl\langle \nabla L_G(\boldsymbol{v}), \boldsymbol{u} \bigr\rangle \\ = f(\boldsymbol{v}+\boldsymbol{u}) - f(\boldsymbol{u}) - \bigl\langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \bigr\rangle.$$

Therefore, it suffices to bound the ratio

$$\rho(\boldsymbol{v}) \stackrel{\text{def}}{=} \frac{\int_{\mathcal{U}^c} \exp\bigl\{ f(\boldsymbol{v}+\boldsymbol{u}) - f(\boldsymbol{u}) - \bigl\langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \bigr\rangle \bigr\} d\boldsymbol{u}}{\int_{\mathcal{U}} \exp\bigl\{ f(\boldsymbol{v}+\boldsymbol{u}) - f(\boldsymbol{u}) - \bigl\langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \bigr\rangle \bigr\} d\boldsymbol{u}}$$

for the elliptic set $\mathcal{U} = \mathcal{U}(\boldsymbol{v}, \mathtt{r}_0) = \bigl\{ \boldsymbol{u} \colon \| D(\boldsymbol{v}) \boldsymbol{u} \| \le \mathtt{r}_0 \bigr\}$ uniformly in $\boldsymbol{v}$ from the set $\bigl\{ \boldsymbol{v} \colon \bigl\| D_G\bigl(\boldsymbol{v} - \boldsymbol{v}_G^*\bigr) \bigr\| \le \mathtt{r}_G \bigr\}$; see Theorem 31.

First we present some bounds for the denominator of $\rho(\boldsymbol{v})$. Local smoothness of $f(\boldsymbol{v}) = \mathbb{E} L_G(\boldsymbol{v})$ implies

$$\int_{\mathcal{U}} \exp\{f(\boldsymbol{v}+\boldsymbol{u}) - f(\boldsymbol{u}) - \langle \nabla f(\boldsymbol{v}), \boldsymbol{u}\rangle\} \, d\boldsymbol{u}$$
$$\approx \int_{\mathcal{U}} \exp\Bigl(-\frac{\|D_G(\boldsymbol{v})\boldsymbol{u}\|^2}{2}\Bigr) \, d\boldsymbol{u},$$

Moreover,

$$\frac{\det D_G(\boldsymbol{v})}{(2\pi)^{p/2}} \int_{\mathcal{U}} \exp\Bigl(-\frac{\|D_G(\boldsymbol{v})\boldsymbol{u}\|^2}{2}\Bigr) \, d\boldsymbol{u} = \mathbb{P}\bigl(\bigl\|D(\boldsymbol{v}) D_G^{-1}(\boldsymbol{v})\boldsymbol{\gamma}\bigr\| \le \mathtt{r}_0\bigr)$$

for $\boldsymbol{\gamma} \sim \mathcal{N}(0, I_p)$. The choice $\mathtt{r}_0 \ge \sqrt{\mathtt{p}_G(\boldsymbol{v})} + \sqrt{2\mathtt{x}}$ yields

$$\mathbb{P}\bigl(\bigl\|D(\boldsymbol{v}) D_G^{-1}(\boldsymbol{v})\boldsymbol{\gamma}\bigr\| \le \mathtt{r}_0\bigr) \ge 1 - \mathrm{e}^{-\mathtt{x}}.$$

$f(\boldsymbol{v}) = \mathbb{E} L(\boldsymbol{v})$ is concave and $-\bigl\langle \nabla^2 f(\boldsymbol{v}) \boldsymbol{u}, \boldsymbol{u} \bigr\rangle = \| D(\boldsymbol{v}) \boldsymbol{u} \|^2$.
For any $\boldsymbol{u}$ with $\| D(\boldsymbol{v}) \boldsymbol{u} \| = \mathtt{r} > \mathtt{r}_0$

$$
\begin{aligned}
& f(\boldsymbol{v}+\boldsymbol{u}) - f(\boldsymbol{v}) - \bigl\langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \bigr\rangle - \| G \boldsymbol{u} \|^2 / 2 \\
& \qquad \le -\mathtt{C}_0 \bigl(\| D(\boldsymbol{v}) \boldsymbol{u} \| \mathtt{r}_0 - \mathtt{r}_0^2 / 2 \bigr) - \| G \boldsymbol{u} \|^2 / 2 \\
& \qquad = -\mathtt{C}_0 \bigl(\| D(\boldsymbol{v}) \boldsymbol{u} \| \mathtt{r}_0 - \mathtt{r}_0^2 / 2 \bigr) - \| D_G(\boldsymbol{v}) \boldsymbol{u} \|^2 / 2 + \| D(\boldsymbol{v}) \boldsymbol{u} \|^2 / 2.
\end{aligned}
$$

with $\mathtt{C}_0 \ge 1/2$ and $D_G^2(\boldsymbol{v}) = D^2(\boldsymbol{v}) + G^2$.
Now we can use the result about Gaussian integrals:

$$
\begin{aligned}
& \frac{\det D_G}{(2\pi)^{p/2}} \int_{\| D \boldsymbol{u} \| \ge \mathtt{r}_0} \exp\Bigl\{ -\Bigl(\| D \boldsymbol{u} \| \mathtt{r}_0 - \mathtt{r}_0^2 / 2 - \| D_G \boldsymbol{u} \|^2 + \| D \boldsymbol{u} \|^2 \Bigr) / 2 \Bigr\} \, d\boldsymbol{u} \\
& \quad \le \mathtt{C} \mathrm{e}^{-(\mathtt{p}_G(\boldsymbol{v}) + \mathtt{x})/2}.
\end{aligned}
$$

- van der Vaart, A. W. , van Zanten, J. H. (2008)
- Leahu, H. (2011). Gaussian models
- Castillo, I. , Nickl, R. (2013, 2014). General results.
- Panov and Sp (2015). Semiparametric problems

Contraction results and local quadratic approximation of the log-likelihood. The Gaussian approximation in a weak sense.

Theorem Suppose that It holds on $\Omega(\mathtt{x})$ with $\widetilde{D}_G^2 = D_G^2(\widetilde{\boldsymbol{\upsilon}}_G)$

$$\sup_{A \in \mathcal{B}_s(\mathbb{R}^p)} \left| \mathbb{P}\bigl(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \in A \bigm| \boldsymbol{Y}\bigr) - \mathbb{P}'\bigl(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\bigr) \right| \leq \mathtt{C}\diamondsuit(\mathtt{r}_0)$$

$$\sup_{A \in \mathcal{B}(\mathbb{R}^p)} \left| \mathbb{P}\bigl(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \in A \bigm| \boldsymbol{Y}\bigr) - \mathbb{P}'\bigl(\widetilde{D}_G^{-1}\boldsymbol{\gamma} \in A\bigr) \right| \leq \mathtt{C}\,\delta_3(\mathtt{r}_0)$$

with

$$\delta_3(\mathtt{r}_0) \lesssim \frac{\mathtt{r}_0^3}{\sqrt{n}} \lesssim \frac{\mathtt{p}_G^{3/2}}{\sqrt{n}},$$

$$\diamondsuit(\mathtt{r}_0) \lesssim \frac{\mathtt{r}_0^6}{n} \lesssim \frac{\mathtt{p}_G^3}{n}.$$

Fix any centrally symmetric set A. First we restrict the posterior probability to the set $\widetilde{\mathcal{A}}(\mathtt{r}_0) = \{\boldsymbol{u} \colon \|\widetilde{D}\boldsymbol{u}\| \le \mathtt{r}_0\}$. Then we apply the quadratic approximation of the log-likelihood function $L(\boldsymbol{\upsilon})$. Denote $A(\mathtt{r}_0) = A \cap \widetilde{\mathcal{A}}(\mathtt{r}_0)$. Obviously, $A(\mathtt{r}_0)$ is centrally symmetric as well. Further,

$$\boldsymbol{I\!P}\bigl(\boldsymbol{\upsilon}_G - \widetilde{\boldsymbol{\upsilon}}_G \in A \,\big|\, \boldsymbol{Y}\bigr) = \frac{\int_A \exp\bigl\{L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u})\bigr\} d\boldsymbol{u}}{\int_{I\!R^p} \exp\bigl\{L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u})\bigr\} d\boldsymbol{u}}$$

$$\le \frac{\int_{A(\mathtt{r}_0)} \exp\bigl\{L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{\upsilon}}_G), \boldsymbol{u} \bigr\rangle\bigr\} d\boldsymbol{u}}{\int_{\widetilde{\mathcal{A}}(\mathtt{r}_0)} \exp\bigl\{L_G(\widetilde{\boldsymbol{\upsilon}}_G + \boldsymbol{u}) - L_G(\widetilde{\boldsymbol{\upsilon}}_G) - \bigl\langle \nabla L_G(\widetilde{\boldsymbol{\upsilon}}_G), \boldsymbol{u} \bigr\rangle\bigr\} d\boldsymbol{u}} + \rho(\mathtt{r}_0).$$

Fix $\boldsymbol{\upsilon} \in \mathcal{A}_G(\mathtt{r}_G)$. Then for $A \subset \mathcal{U}$ centrally symmetric

$$\int_A \exp\bigl\{ f(\boldsymbol{\upsilon} + \boldsymbol{u}) - f(\boldsymbol{u}) - \bigl\langle \nabla f(\boldsymbol{\upsilon}), \boldsymbol{u} \bigr\rangle \bigr\} \, d\boldsymbol{u}$$

$$\geq \bigl(1 - \diamondsuit(\mathtt{r}_0)\bigr) \int_A \exp\Bigl(- \frac{\|D_G(\boldsymbol{\upsilon})\boldsymbol{u}\|^2}{2} \Bigr) \, d\boldsymbol{u},$$

$$\int_A \exp\bigl\{ f(\boldsymbol{\upsilon} + \boldsymbol{u}) - f(\boldsymbol{u}) - \bigl\langle \nabla f(\boldsymbol{\upsilon}), \boldsymbol{u} \bigr\rangle \bigr\} \, d\boldsymbol{u}$$

$$\leq \bigl(1 + \diamondsuit(\mathtt{r}_0)\bigr) \int_A \exp\Bigl(- \frac{\|D_G(\boldsymbol{\upsilon})\boldsymbol{u}\|^2}{2} \Bigr) \, d\boldsymbol{u},$$

where $\diamondsuit(\mathtt{r}_0) = 4\delta_3^2 + 4\delta_4$.

Abraham, K. and Nickl, R. (2019).
On statistical Calderón problems.

Asi, H. and Duchi, J. C. (2019).
Stochastic (Approximate) Proximal Point Methods: Convergence, Optimality, and Adaptivity.
SIAM Journal on Optimization, 29(3):2257–2290.

Nesterov, Y. and Spokoiny, V. (2017).
Random gradient-free minimization of convex functions.
Foundations of Computational Mathematics, 17(2):527–566.

Nickl, R. (2017).
Bernstein - von Mises theorems for statistical inverse problems I: Schrödinger equation.
arXiv e-prints, page arXiv:1707.01764.

Nickl, R. and Söhl, J. (2019).
Bernstein – von Mises theorems for statistical inverse problems II: compound Poisson processes.
Electronic Journal of Statistics, 13(2):3513–3571.

Nickl, R., van de Geer, S. A., and Wang, S. (2018).
Convergence rates for Penalised Least Squares Estimators in PDE-constrained regression problems.

References

Spokoiny, V. and Panov, M. (2019).
Accuracy of Gaussian approximation in nonparametric Bernstein – von Mises Theorem.
preprint arXiv:https:1910.06028.

Trabs, M. (2018).
Bayesian inverse problems with unknown operators.
Inverse Problems, 34(8):085001.