Weierstraß-Institut für
Angewandte Analysis und Stochastik

Leibniz
Gemeinschaft

Linearly pertubed optimization: theory and applications

Vladimir Spokoiny ,
WIAS, HU Berlin

7. Oktober 2024

1 Introduction

2 Linearly pertubed optimization

- Quadratic case
- 2S expansions
- Linear perturbation under third order smoothness
- Uniform smoothness

3 Fourth order approximation

4 Quadratic penalization

Let $f(\boldsymbol{\upsilon})$ be a smooth concave function,

$$\boldsymbol{\upsilon}^* = \operatorname*{argmax}_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon}), \quad \mathbb{F} = -\nabla^2 f(\boldsymbol{\upsilon}^*).$$

Let another function $g(\boldsymbol{\upsilon})$ satisfy for some vector $\boldsymbol{A}$

$$g(\boldsymbol{\upsilon}) - g(\boldsymbol{\upsilon}^*) = \langle \boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*, \boldsymbol{A} \rangle + f(\boldsymbol{\upsilon}) - f(\boldsymbol{\upsilon}^*). \tag{1}$$

Define

$$\boldsymbol{\upsilon}^\circ \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon}), \qquad g(\boldsymbol{\upsilon}^\circ) = \max_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon}). \tag{2}$$

Aim: evaluate the quantities $\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*$ and $g(\boldsymbol{\upsilon}^\circ) - g(\boldsymbol{\upsilon}^*)$.

Let $L(\boldsymbol{\upsilon})$ be a log-likelihood function. Consider the MLE

$$\widetilde{\boldsymbol{\upsilon}} = \operatorname*{argmax}_{\boldsymbol{\upsilon}} L(\boldsymbol{\upsilon})$$

and the background truth

$$\boldsymbol{\upsilon}^{*} = \operatorname*{argmax}_{\boldsymbol{\upsilon}} \mathbb{E} L(\boldsymbol{\upsilon}).$$

Stochastically linear smooth (SLS) models: $\mathbb{E} L(\boldsymbol{\upsilon})$ is smooth and concave in $\boldsymbol{\upsilon}$ and $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E} L(\boldsymbol{\upsilon})$ is linear in $\boldsymbol{\upsilon}$:

$$\boldsymbol{A} = \nabla \zeta(\boldsymbol{\upsilon}) = \nabla \zeta.$$

Outcome: Fisher theorem and Wilks phenomenon in statistics.

Let $h(\cdot)$ be concave and

$$\boldsymbol{v}^* = \operatorname{argmax} h(\boldsymbol{v}).$$

Consider

$$\begin{aligned} g(\boldsymbol{v}) &= h(\boldsymbol{v}) - \|G\boldsymbol{v}\|^2/2, \\ f(\boldsymbol{v}) &= h(\boldsymbol{v}) - \|G\boldsymbol{v}\|^2/2 + \langle G^2\boldsymbol{v}^*, \boldsymbol{v}\rangle, \quad . \end{aligned}$$

Then $\nabla f(\boldsymbol{v}^*) = 0$ and $\boldsymbol{v}^* = \operatorname{argmax} f(\boldsymbol{v})$.

g is a linear perturbation of f with $\boldsymbol{A} = -G^2\boldsymbol{v}^*$.

Outcome: roughness penalty, effective dimension, critical dimension.

Let f be a concave function and

$$\boldsymbol{v}^* = \operatorname{argmax} f(\boldsymbol{v}).$$

Let also $\boldsymbol{v}^\circ$ be a current guess. Define

$$g(\boldsymbol{v}) = f(\boldsymbol{v}) - \langle \nabla f(\boldsymbol{v}^\circ), \boldsymbol{v} - \boldsymbol{v}^\circ \rangle.$$

Then $\nabla g(\boldsymbol{v}^\circ) = 0$ and hence,

$$\boldsymbol{v}^\circ = \operatorname{argmax} g(\boldsymbol{v}).$$

g is a linear perturbation of f with $\boldsymbol{A} = \nabla f(\boldsymbol{v}^\circ)$.

Outcome: Newton – Kantorovich – Nemirovskii-Nesterov theorem on quadratic convergence of strongly convex optimization.

Let $\mathbb{P}_f \sim \exp f(\boldsymbol{x})$. Denote by $\mathbb{N}_{\boldsymbol{x},\mathbb{Z}}$ the Gaussian measure with the mean $\boldsymbol{x}$ and covariance $\mathbb{Z}^{-1}$, i.e. $\mathbb{N}_{\boldsymbol{x},\mathbb{Z}} \stackrel{\text{def}}{=} \mathcal{N}(\boldsymbol{x}, \mathbb{Z}^{-1})$.

$$\text{Gauss VI:} \quad (\boldsymbol{x}_{\text{VI}}, \mathbb{Z}_{\text{VI}}) = \operatorname*{arginf}_{\boldsymbol{x},\mathbb{Z}} \mathscr{K}(\mathbb{N}_{\boldsymbol{x},\mathbb{Z}} \,\|\, \mathbb{P}_f).$$

Natural candidates:

1. Laplace: $\boldsymbol{x}_{\text{VI}} \approx \operatorname{argmax} f(\boldsymbol{x})$, $\mathbb{Z}_{\text{VI}} \approx -\nabla^2 f(\boldsymbol{x}^*)$;
2. Moments: $\boldsymbol{x}_{\text{VI}} \approx \mathbb{E}_f \boldsymbol{X}$, $\mathbb{Z}_{\text{VI}}^{-1} \approx \operatorname{Var}_f(\boldsymbol{X})$.

Let $\mathbb{P}_f \sim \exp f(\boldsymbol{x})$. Denote by $\mathbb{N}_{\boldsymbol{x},\mathbb{Z}}$ the Gaussian measure with the mean $\boldsymbol{x}$ and covariance $\mathbb{Z}^{-1}$, i.e. $\mathbb{N}_{\boldsymbol{x},\mathbb{Z}} \stackrel{\text{def}}{=} \mathcal{N}(\boldsymbol{x}, \mathbb{Z}^{-1})$.

$$\text{Gauss VI:} \quad (\boldsymbol{x}_{\text{VI}}, \mathbb{Z}_{\text{VI}}) = \operatorname*{arginf}_{\boldsymbol{x},\mathbb{Z}} \mathscr{K}(\mathbb{N}_{\boldsymbol{x},\mathbb{Z}} \,\|\, \mathbb{P}_f).$$

Natural candidates:

1. Laplace: $\boldsymbol{x}_{\text{VI}} \approx \operatorname{argmax} f(\boldsymbol{x})$, $\mathbb{Z}_{\text{VI}} \approx -\nabla^2 f(\boldsymbol{x}^*)$;
2. Moments: $\boldsymbol{x}_{\text{VI}} \approx \mathbb{E}_f \boldsymbol{X}$, $\mathbb{Z}_{\text{VI}}^{-1} \approx \operatorname{Var}_f(\boldsymbol{X})$.

[Katsevich and Rigollet, 2023] argued for (2).

Lemma

Let $f(\boldsymbol{v})$ *be quadratic with* $\nabla^2 f(\boldsymbol{v}) \equiv -\mathbb{F}$ *. If* $g(\boldsymbol{v})$ *satisfy* (1), *then*

$$\boldsymbol{v}^\circ - \boldsymbol{v}^* = \mathbb{F}^{-1}\boldsymbol{A}, \qquad g(\boldsymbol{v}^\circ) - g(\boldsymbol{v}^*) = \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2.$$

Proof. Clearly $-\nabla^2 g(\boldsymbol{v}) \equiv -\mathbb{F}$ and

$$\nabla g(\boldsymbol{v}^*) - \nabla g(\boldsymbol{v}^\circ) = \mathbb{F}(\boldsymbol{v}^\circ - \boldsymbol{v}^*).$$

Further, (1) and $\nabla f(\boldsymbol{v}^*) = 0$ yield $\nabla g(\boldsymbol{v}^*) = \boldsymbol{A}$. Together with $\nabla g(\boldsymbol{v}^\circ) = 0$, this implies $\boldsymbol{v}^\circ - \boldsymbol{v}^* = \mathbb{F}^{-1}\boldsymbol{A}$.
Taylor expansion of g at $\boldsymbol{v}^\circ$ yields by $\nabla g(\boldsymbol{v}^\circ) = 0$

$$g(\boldsymbol{v}^*) - g(\boldsymbol{v}^\circ) = -\frac{1}{2}(\boldsymbol{v}^\circ - \boldsymbol{v}^*)^\top \mathbb{F}(\boldsymbol{v}^\circ - \boldsymbol{v}^*) = -\frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 .$$

Define

$$\delta_3(\boldsymbol{v},\boldsymbol{u}) = f(\boldsymbol{v}+\boldsymbol{u}) - f(\boldsymbol{v}) - \langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \rangle - \frac{1}{2}\langle \nabla^2 f(\boldsymbol{v}), \boldsymbol{u}^{\otimes 2} \rangle,$$

$$\delta_3'(\boldsymbol{v},\boldsymbol{u}) = \langle \nabla f(\boldsymbol{v}+\boldsymbol{u}), \boldsymbol{u} \rangle - \langle \nabla f(\boldsymbol{v}), \boldsymbol{u} \rangle - \langle \nabla^2 f(\boldsymbol{v}), \boldsymbol{u}^{\otimes 2} \rangle .$$

For $\mathbb{D}^2 \le \mathbb{F}(\boldsymbol{v}) = -\nabla^2 f(\boldsymbol{v})$, define

$$\omega(\boldsymbol{v}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{u}:\, \|\mathbb{D}\boldsymbol{u}\| \le \mathtt{r}} \frac{2|\delta_3(\boldsymbol{v},\boldsymbol{u})|}{\|\mathbb{D}\boldsymbol{u}\|^2} ,$$

$$\omega'(\boldsymbol{v}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{u}:\, \|\mathbb{D}\boldsymbol{u}\| \le \mathtt{r}} \frac{|\delta_3'(\boldsymbol{v},\boldsymbol{u})|}{\|\mathbb{D}\boldsymbol{u}\|^2} . \tag{3}$$

Proposition

Fix $\nu \leq 2/3$ *and* $\mathtt{r}$ *such that* $\|\mathbb{F}^{-1/2}\boldsymbol{A}\| \leq \nu\,\mathtt{r}$ *. Suppose now that* $f(\boldsymbol{\upsilon})$ *satisfy* (3) *for* $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}^{*}$ *,* $\mathbb{D} = \mathbb{F}^{1/2}$ *, and* ω' *such that*

$$1 - \nu - \omega' \geq 0. \tag{4}$$

Then for $\boldsymbol{\upsilon}^{\circ}$ *from* (2)*, it holds*

$$\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\| \leq \mathtt{r}\,.$$

With $\mathbb{D} = \mathbb{F}^{1/2}$, the bound $\|\mathbb{D}^{-1}\boldsymbol{A}\| \le \nu\, \mathtt{r}$ implies for any $\boldsymbol{u}$

$$\bigl|\langle \boldsymbol{A}, \boldsymbol{u} \rangle\bigr| = \bigl|\langle \mathbb{D}^{-1}\boldsymbol{A}, \mathbb{D}\boldsymbol{u} \rangle\bigr| \le \nu\, \mathtt{r} \|\mathbb{D}\boldsymbol{u}\| .$$

If $\|\mathbb{D}\boldsymbol{u}\| > \mathtt{r}$, then $\mathtt{r}\|\mathbb{D}\boldsymbol{u}\| \le \|\mathbb{D}\boldsymbol{u}\|^2$. Therefore,

$$\bigl|\langle \boldsymbol{A}, \boldsymbol{u} \rangle\bigr| \le \nu \|\mathbb{D}\boldsymbol{u}\|^2 , \qquad \|\mathbb{D}\boldsymbol{u}\| > \mathtt{r} . \tag{5}$$

Let $\boldsymbol{v}$ satisfy $\|\mathbb{D}(\boldsymbol{v} - \boldsymbol{v}^*)\| = \mathtt{r}$. Denote $\boldsymbol{u} = \boldsymbol{v} - \boldsymbol{v}^*$. The idea is to show that the derivative $\frac{d}{dt} g(\boldsymbol{v}^* + t\boldsymbol{u}) < 0$ is negative for $t > 1$. Then all the extreme points of $g(\boldsymbol{v})$ are within $\mathcal{A}(\mathtt{r})$. We use the decomposition

$$g(\boldsymbol{v}^* + t\boldsymbol{u}) - g(\boldsymbol{v}^*) = \langle \boldsymbol{A}, \boldsymbol{u} \rangle\, t + f(\boldsymbol{v}^* + t\boldsymbol{u}) - f(\boldsymbol{v}^*).$$

With $h(t) = f(\boldsymbol{v}^* + t\boldsymbol{u}) - f(\boldsymbol{v}^*) - \langle \boldsymbol{A}, \boldsymbol{u} \rangle\, t$, it holds

$$\frac{d}{dt} f(\boldsymbol{v}^* + t\boldsymbol{u}) = \langle \boldsymbol{A}, \boldsymbol{u} \rangle + h'(t). \tag{6}$$

By definition of $\boldsymbol{v}^*$, it also holds $h'(0) = -\langle \boldsymbol{A}, \boldsymbol{u} \rangle$. The identity $\nabla^2 f(\boldsymbol{v}^*) = -\mathbb{D}^2$ yields $h''(0) = -\|\mathbb{D}\boldsymbol{u}\|^2$. Bound (3) implies for $|t| \leq 1$

$$\big| h'(t) - h'(0) - t h''(0) \big| \leq t^2 \big| h''(0) \big| \, \omega' \, .$$

For $t = 1$, we obtain by (4) and (5)

$$h'(1) \leq -\langle \boldsymbol{A}, \boldsymbol{u} \rangle + h''(0) - h''(0) \, \omega' \leq -\big| h''(0) \big| (1 - \omega' - \nu) < 0.$$

Moreover, concavity of $h(t)$ imply that $h'(t) - h'(0)$ decreases in t for $t > 1$. Further, summing up the above derivation yields

$$\frac{d}{dt} h(\boldsymbol{v}^* + t\boldsymbol{u}) \Big|_{t=1} \leq -\|\mathbb{D}\boldsymbol{u}\|^2 (1 - \nu - \omega') < 0.$$

As $\frac{d}{dt} h(\boldsymbol{v}^* + t\boldsymbol{u})$ decreases with $t \geq 1$ together with $h'(t)$ due to (6), the same applies to all such t. This implies the assertion.

Proposition

Under the conditions of Proposition 1, with $\boldsymbol{\xi} = \mathbb{D}^{-1}\boldsymbol{A} = \mathbb{F}^{-1/2}\boldsymbol{A}$

$$-\frac{\omega}{1+\omega}\|\boldsymbol{\xi}\|^2 \leq 2g(\boldsymbol{\upsilon}^{\circ}) - 2g(\boldsymbol{\upsilon}^{*}) - \|\boldsymbol{\xi}\|^2 \leq \frac{\omega}{1-\omega}\|\boldsymbol{\xi}\|^2 . \tag{7}$$

Also

$$\begin{aligned} \|\mathbb{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*} - \mathbb{F}^{-1}\boldsymbol{A})\|^2 &\leq \frac{3\omega}{(1-\omega)^2}\,\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 , \\ \|\mathbb{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\| &\leq \frac{1+\sqrt{2\omega}}{1-\omega}\,\|\mathbb{F}^{-1/2}\boldsymbol{A}\| . \end{aligned} \tag{8}$$

Proof

By (3), for any $\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})$

$$\left| f(\boldsymbol{\upsilon}^*) - f(\boldsymbol{\upsilon}) - \frac{1}{2} \|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 \right| \leq \frac{\omega}{2} \|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2. \tag{9}$$

Further,

$$g(\boldsymbol{\upsilon}) - g(\boldsymbol{\upsilon}^*) - \frac{1}{2} \|\mathbb{D}^{-1} \boldsymbol{A}\|^2 = \langle \boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*, \boldsymbol{A} \rangle + f(\boldsymbol{\upsilon}) - f(\boldsymbol{\upsilon}^*) - \frac{1}{2} \|\mathbb{D}^{-1} \boldsymbol{A}\|^2$$

$$= -\frac{1}{2} \left\| \mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \mathbb{D}^{-1} \boldsymbol{A} \right\|^2 + f(\boldsymbol{\upsilon}) - f(\boldsymbol{\upsilon}^*) + \frac{1}{2} \|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2. \tag{10}$$

As $\boldsymbol{\upsilon}^\circ \in \mathcal{A}(\mathtt{r})$ and it maximizes $g(\boldsymbol{\upsilon})$, we derive by (9)

$$g(\boldsymbol{\upsilon}^\circ) - g(\boldsymbol{\upsilon}^*) - \frac{1}{2} \|\mathbb{D}^{-1} \boldsymbol{A}\|^2 = \max_{\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})} \left\{ g(\boldsymbol{\upsilon}) - g(\boldsymbol{\upsilon}^*) - \frac{1}{2} \|\mathbb{D}^{-1} \boldsymbol{A}\|^2 \right\}$$

$$\leq \max_{\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})} \left\{ -\frac{1}{2} \left\| \mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \mathbb{D}^{-1} \boldsymbol{A} \right\|^2 + \frac{\omega}{2} \left\| \mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \right\|^2 \right\}.$$

Further, $\max_{\boldsymbol{u}}\{\omega\|\boldsymbol{u}\|^2 - \|\boldsymbol{u}-\boldsymbol{\xi}\|^2\} = \frac{\omega}{1-\omega}\|\boldsymbol{\xi}\|^2$ for $\omega \in [0,1)$ and $\boldsymbol{\xi} \in I\!R^p$, yielding

$$g(\boldsymbol{v}^\circ) - g(\boldsymbol{v}^*) - \frac{1}{2}\|\mathbb{D}^{-1}\boldsymbol{A}\|^2 \le \frac{\omega}{2(1-\omega)}\|\mathbb{D}^{-1}\boldsymbol{A}\|^2.$$

Similarly

$$\begin{aligned} & g(\boldsymbol{v}^\circ) - g(\boldsymbol{v}^*) - \frac{1}{2}\|\mathbb{D}^{-1}\boldsymbol{A}\|^2 \\ & \ge \max_{\boldsymbol{v}\in\mathcal{A}(\mathtt{r})}\Bigl\{-\frac{1}{2}\bigl\|\mathbb{D}(\boldsymbol{v}-\boldsymbol{v}^*) - \mathbb{D}^{-1}\boldsymbol{A}\bigr\|^2 - \frac{\omega}{2}\bigl\|\mathbb{D}(\boldsymbol{v}-\boldsymbol{v}^*)\bigr\|^2\Bigr\} \\ & = -\frac{\omega}{2(1+\omega)}\,\|\mathbb{D}^{-1}\boldsymbol{A}\|^2. \qquad (11) \end{aligned}$$

These bounds imply (7).

Now we derive similarly to (10) that for $\boldsymbol{v} \in \mathcal{A}(\mathtt{r})$

$$g(\boldsymbol{v}) - g(\boldsymbol{v}^*) \le \langle \boldsymbol{v} - \boldsymbol{v}^*, \boldsymbol{A} \rangle - \frac{1-\omega}{2} \|\mathbb{D}(\boldsymbol{v} - \boldsymbol{v}^*)\|^2.$$

A particular choice $\boldsymbol{v} = \boldsymbol{v}^\circ$ yields

$$g(\boldsymbol{v}^\circ) - g(\boldsymbol{v}^*) \le \langle \boldsymbol{v}^\circ - \boldsymbol{v}^*, \boldsymbol{A} \rangle - \frac{1-\omega}{2} \|\mathbb{D}(\boldsymbol{v}^\circ - \boldsymbol{v}^*)\|^2.$$

Combining this result with (11) allows to bound

$$\langle \boldsymbol{v}^\circ - \boldsymbol{v}^*, \boldsymbol{A} \rangle - \frac{1-\omega}{2} \|\mathbb{D}(\boldsymbol{v}^\circ - \boldsymbol{v}^*)\|^2 - \frac{1}{2} \|\mathbb{D}^{-1} \boldsymbol{A}\|^2 \ge -\frac{\omega}{2(1+\omega)} \|\mathbb{D}^{-1} \boldsymbol{A}\|^2.$$

For $\boldsymbol{\xi} = \mathbb{D}^{-1}\boldsymbol{A}$, $\boldsymbol{u} = \mathbb{D}(\boldsymbol{v}^{\circ} - \boldsymbol{v}^{*})$, and $\omega \in [0, 1/3]$, the inequality

$$2\langle \boldsymbol{u}, \boldsymbol{\xi} \rangle - (1-\omega)\|\boldsymbol{u}\|^2 - \|\boldsymbol{\xi}\|^2 \geq -\frac{\omega}{1+\omega}\|\boldsymbol{\xi}\|^2$$

implies

$$\left\|\boldsymbol{u} - \frac{1}{1-\omega}\boldsymbol{\xi}\right\|^2 \leq \frac{2\omega}{(1+\omega)(1-\omega)^2}\|\boldsymbol{\xi}\|^2$$

yielding for $\omega \leq 1/3$

$$\|\boldsymbol{u} - \boldsymbol{\xi}\| \leq \left(\omega + \sqrt{\frac{2\omega}{1+\omega}}\right)\frac{\|\boldsymbol{\xi}\|}{1-\omega} \leq \frac{\sqrt{3\omega}\|\boldsymbol{\xi}\|}{1-\omega},$$

$$\|\boldsymbol{u}\| \leq \left(1 + \sqrt{\frac{2\omega}{1+\omega}}\right)\frac{\|\boldsymbol{\xi}\|}{1-\omega} \leq \frac{1+\sqrt{2\omega}\|\boldsymbol{\xi}\|}{1-\omega},$$

and (8) follows.

Lemma

It holds

$$\max_{\boldsymbol{u}} \{ \omega \|\boldsymbol{u}\|^2 - \|\boldsymbol{u} - \boldsymbol{\xi}\|^2 \} = \frac{\omega}{1-\omega} \|\boldsymbol{\xi}\|^2 .$$

If $\omega \leq 1/3$ *, then the inequality*

$$2\langle \boldsymbol{u}, \boldsymbol{\xi} \rangle - (1-\omega)\|\boldsymbol{u}\|^2 - \|\boldsymbol{\xi}\|^2 \geq -\frac{\omega}{1+\omega} \|\boldsymbol{\xi}\|^2$$

implies

$$\Bigl\| \boldsymbol{u} - \frac{1}{1-\omega} \boldsymbol{\xi} \Bigr\|^2 \leq \frac{2\omega}{(1+\omega)(1-\omega)^2} \|\boldsymbol{\xi}\|^2$$

$$\|\boldsymbol{u} - \boldsymbol{\xi}\| \leq \left(\omega + \sqrt{\frac{2\omega}{1+\omega}} \right) \frac{\|\boldsymbol{\xi}\|}{1-\omega} \leq \frac{\sqrt{3\omega}\|\boldsymbol{\xi}\|}{1-\omega} .$$

$(\mathcal{T}_3)$ *There exists* τ_3 *such that for all* $\boldsymbol{u}$ *with* $\|\mathbb{D}\boldsymbol{u}\| \leq \mathtt{r}$

$$\left|\delta_3(\boldsymbol{v},\boldsymbol{u})\right| \leq \frac{\tau_3}{6}\|\mathbb{D}\,\boldsymbol{u}\|^3, \quad \left|\delta_3'(\boldsymbol{v},\boldsymbol{u})\right| \leq \frac{\tau_3}{2}\|\mathbb{D}\,\boldsymbol{u}\|^3.$$

$(\mathcal{T}_4)$ *There exists* τ_4 *such that for all* $\boldsymbol{u}$ *with* $\|\mathbb{D}\boldsymbol{u}\| \leq \mathtt{r}$

$$\left|\delta_4(\boldsymbol{v},\boldsymbol{u})\right| \leq \frac{\tau_4}{24}\|\mathbb{D}\,\boldsymbol{u}\|^4.$$

$(\mathcal{T}_3^*)$ $f(\boldsymbol{\upsilon})$ *is strongly concave,* $\mathbb{D}^2 \leq \nabla^2 f(\boldsymbol{\upsilon})$ *, and*

$$\sup_{\boldsymbol{u}\colon \|\mathbb{D}\boldsymbol{u}\|\leq \mathtt{r}} \sup_{\boldsymbol{z}\in\mathbb{R}^p} \frac{\bigl|\langle \nabla^3 f(\boldsymbol{\upsilon}+\boldsymbol{u}), \boldsymbol{z}^{\otimes 3}\rangle\bigr|}{\|\mathbb{D}\boldsymbol{z}\|^3} \leq \tau_3 .$$

$(\mathcal{T}_4^*)$ $f(\boldsymbol{\upsilon})$ *is strongly concave,* $\mathbb{D}^2 \leq \nabla^2 f(\boldsymbol{\upsilon})$ *, and*

$$\sup_{\boldsymbol{u}\colon \|\mathbb{D}\boldsymbol{u}\|\leq \mathtt{r}} \sup_{\boldsymbol{z}\in\mathbb{R}^p} \frac{\bigl|\langle \nabla^4 f(\boldsymbol{\upsilon}+\boldsymbol{u}), \boldsymbol{z}^{\otimes 4}\rangle\bigr|}{\|\mathbb{D}\boldsymbol{z}\|^4} \leq \tau_4 .$$

Banach's characterization [Banach, 1938] yields under $(\mathcal{T}_3^*)$ (resp $(\mathcal{T}_4^*)$)

$$\bigl|\langle \nabla^3 f(\boldsymbol{\upsilon}+\boldsymbol{u}), \boldsymbol{z}_1 \otimes \boldsymbol{z}_2 \otimes \boldsymbol{z}_3\rangle\bigr| \leq \tau_3 \|\mathbb{D}\boldsymbol{z}_1\| \, \|\mathbb{D}\boldsymbol{z}_2\| \, \|\mathbb{D}\boldsymbol{z}_3\| \, ; \tag{12}$$

$$\bigl|\langle \nabla^4 f(\boldsymbol{\upsilon}+\boldsymbol{u}), \boldsymbol{z}_1 \otimes \boldsymbol{z}_2 \otimes \boldsymbol{z}_3 \otimes \boldsymbol{z}_4\rangle\bigr| \leq \tau_4 \prod_{k=1}^{4} \|\mathbb{D}\boldsymbol{z}_k\| \, . \tag{13}$$

Proposition

Let $f(\boldsymbol{\upsilon})$ *be a strongly concave function with* $f(\boldsymbol{\upsilon}^{*}) = \max_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon})$ *and* $\mathbb{F} = -\nabla^2 f(\boldsymbol{\upsilon}^{*})$ *. Let* $g(\boldsymbol{\upsilon})$ *fulfill* (1) *with some vector* $\boldsymbol{A}$ *. Suppose that* $f(\boldsymbol{\upsilon})$ *follows* $(\boldsymbol{\mathcal{T}_3})$ *with* $\mathtt{r} = \nu^{-1}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|$ *for* $\nu < 1$ *and some* $\tau_3 \geq 0$ *. Let*

$$\tau_3 \, \|\mathbb{F}^{-1/2}\boldsymbol{A}\| < 2\nu(1-\nu).$$

Then $\boldsymbol{\upsilon}^{\circ} = \operatorname{argmax}_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon})$ *satisfies*

$$\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\| \leq \nu^{-1}\|\mathbb{F}^{-1/2}\boldsymbol{A}\| \, .$$

Proposition

Under the conditions of Proposition 3

$$
\begin{aligned}
-\frac{2\tau_3}{3}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^3 &\leq 2g(\boldsymbol{\upsilon}^\circ) - 2g(\boldsymbol{\upsilon}^*) - \|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 \\
&\leq \tau_3\,\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^3\,.
\end{aligned} \tag{14}
$$

Moreover, under $(\boldsymbol{\mathcal{T}_3^*})$

$$
\begin{aligned}
\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*) - \mathbb{F}^{-1/2}\boldsymbol{A}\| &\leq \frac{3\tau_3}{4}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2\,, \\
\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\| &\leq \|\mathbb{F}^{-1/2}\boldsymbol{A}\| + \frac{3\tau_3}{4}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2\,.
\end{aligned} \tag{15}
$$

By $(\mathcal{T}_3)$ and $\nabla f(\boldsymbol{\upsilon}^*) = 0$, for any $\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})$

$$\left| f(\boldsymbol{\upsilon}^*) - f(\boldsymbol{\upsilon}) - \frac{1}{2}\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2 \right| \le \frac{\tau_3}{6}\|\mathbb{D}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^3$$
$$\le \frac{\tau_3}{6}\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^3. \qquad (16)$$

Further,

$$g(\boldsymbol{\upsilon}) - g(\boldsymbol{\upsilon}^*) - \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2$$
$$= \langle \boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*, \boldsymbol{A} \rangle + f(\boldsymbol{\upsilon}) - f(\boldsymbol{\upsilon}^*) - \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2$$
$$= -\frac{1}{2}\left\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) - \mathbb{F}^{-1/2}\boldsymbol{A}\right\|^2 + f(\boldsymbol{\upsilon}) - f(\boldsymbol{\upsilon}^*) + \frac{1}{2}\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\|^2.$$

As $\boldsymbol{\upsilon}^{\circ} \in \mathcal{A}(\mathtt{r})$ and it maximizes $g(\boldsymbol{\upsilon})$, we derive by (16) and Lemma 3

$$g(\boldsymbol{\upsilon}^{\circ}) - g(\boldsymbol{\upsilon}^{*}) - \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 = \max_{\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})} \Bigl\{ g(\boldsymbol{\upsilon}) - g(\boldsymbol{\upsilon}^{*}) - \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 \Bigr\}$$

$$\leq \max_{\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})} \Bigl\{ -\frac{1}{2}\bigl\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*}) - \mathbb{F}^{-1/2}\boldsymbol{A}\bigr\|^2 + \frac{\tau_3}{6}\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*})\|^3 \Bigr\}$$

$$\leq \frac{\tau_3}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^3 .$$

Now (14) follows from this and

$$g(\boldsymbol{\upsilon}^{\circ}) - g(\boldsymbol{\upsilon}^{*}) - \frac{1}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2$$

$$\geq \max_{\boldsymbol{\upsilon} \in \mathcal{A}(\mathtt{r})} \Bigl\{ -\frac{1}{2}\bigl\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*}) - \mathbb{F}^{-1/2}\boldsymbol{A}\bigr\|^2 - \frac{\tau_3}{6}\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*})\|^3 \Bigr\}$$

$$\geq -\frac{\tau_3}{3}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^3 .$$

For proving (21) use that $\nabla f(\boldsymbol{v}^*) = 0$, $\nabla g(\boldsymbol{v}^\circ) = 0$, $\nabla f(\boldsymbol{v}^\circ) = \nabla g(\boldsymbol{v}^\circ) - \boldsymbol{A} = -\boldsymbol{A}$, and $-\nabla^2 f(\boldsymbol{v}^*) = \mathbb{F}$. By Lemma 4 with $\boldsymbol{u} = \mathbb{F}^{-1}\boldsymbol{A}$

$$\left\|\mathbb{F}^{-1/2}\{\nabla f(\boldsymbol{v}^* + \mathbb{F}^{-1}\boldsymbol{A}) + \boldsymbol{A}\}\right\| \le \frac{\tau_3}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 . \qquad (17)$$

Further, by (1)

$$\left\|\mathbb{F}^{-1/2}\nabla g(\boldsymbol{v}^* + \mathbb{F}^{-1}\boldsymbol{A})\right\| = \left\|\mathbb{F}^{-1/2}\{\nabla g(\boldsymbol{v}^* + \mathbb{F}^{-1}\boldsymbol{A}) - \boldsymbol{A} + \boldsymbol{A}\}\right\|$$

$$\le \left\|\mathbb{F}^{-1/2}\{\nabla f(\boldsymbol{v}^* + \mathbb{F}^{-1}\boldsymbol{A}) + \boldsymbol{A}\}\right\| \le \frac{\tau_3}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2.$$

By definition $\nabla g(\boldsymbol{v}^\circ) = 0$. This yields

$$\|\mathbb{F}^{-1/2}\{\nabla g(\boldsymbol{v}^* + \mathbb{F}^{-1}\boldsymbol{A}) - \nabla g(\boldsymbol{v}^\circ)\}\| \le \frac{\tau_3}{2}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 . \qquad (18)$$

Now we can use with $\varDelta = \boldsymbol{\upsilon}^* + \mathbb{F}^{-1}\boldsymbol{A} - \boldsymbol{\upsilon}^\circ$

$$\begin{aligned}&\mathbb{F}^{-1/2}\{\nabla g(\boldsymbol{\upsilon}^* + \mathbb{F}^{-1}\boldsymbol{A}) - \nabla g(\boldsymbol{\upsilon}^\circ)\}\\ &\quad = \biggl(\int_0^1 \mathbb{F}^{-1/2}\, \nabla^2 g(\boldsymbol{\upsilon}^\circ + t\varDelta)\, \mathbb{F}^{-1/2}\, dt\biggr)\, \mathbb{F}^{1/2}\varDelta\,.\end{aligned}$$

By (1) $\nabla^2 g(\boldsymbol{\upsilon}) = \nabla^2 f(\boldsymbol{\upsilon})$ for all $\boldsymbol{\upsilon}$. If $\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}$, then $(\boldsymbol{\mathcal{T}_3^*})$ implies $\|\mathbb{F}^{-1/2}\,\nabla^2 f(\boldsymbol{\upsilon})\,\mathbb{F}^{-1/2} + I\!\!I_p\| \le \omega^+ \le \tau_3\,\mathtt{r} \le 1/3$. Hence,

$$\|\mathbb{F}^{-1/2}\{\nabla g(\boldsymbol{\upsilon}^* + \mathbb{F}^{-1}\boldsymbol{A}) - \nabla g(\boldsymbol{\upsilon}^\circ)\}\| \ge (1-\omega^+)\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^* - \mathbb{F}^{-1}\boldsymbol{A})\|\,.$$

This and (26) yield

$$\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^* - \mathbb{F}^{-1}\boldsymbol{A})\| \le \frac{\tau_3}{2(1-\omega^+)}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 \le \frac{3\tau_3}{4}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2\,,$$

and (21) follows.

Lemma

For any $\boldsymbol{\xi} \in \mathbb{R}^p$ *with* $\|\boldsymbol{\xi}\| \leq 2\mathtt{r}/3$ *and* τ *with* $\tau\,\mathtt{r} \leq 1/2$, *it holds*

$$\max_{\|\boldsymbol{u}\|\leq\mathtt{r}} \left(\frac{\tau}{3}\|\boldsymbol{u}\|^3 - \|\boldsymbol{u}-\boldsymbol{\xi}\|^2 \right) \leq \frac{\tau}{2}\,\|\boldsymbol{\xi}\|^3, \tag{19}$$

$$\min_{\|\boldsymbol{u}\|\leq\mathtt{r}} \left(\frac{\tau}{3}\|\boldsymbol{u}\|^3 + \|\boldsymbol{u}-\boldsymbol{\xi}\|^2 \right) \leq \frac{\tau}{3}\,\|\boldsymbol{\xi}\|^3. \tag{20}$$

Any maximizer $\boldsymbol{u}$ of the left hand-side of (19) satisfies

$$\tau\|\boldsymbol{u}\|^{1/2}\boldsymbol{u} - 2(\boldsymbol{u}-\boldsymbol{\xi}) = 0\,.$$

Therefore, $\boldsymbol{u} = \rho\boldsymbol{\xi}$ for some ρ, reducing the problem to the univariate case:

$$\max_{\|\boldsymbol{u}\|\le \mathtt{r}} \Bigl(\frac{\tau}{3}\|\boldsymbol{u}\|^3 - \|\boldsymbol{u}-\boldsymbol{\xi}\|^2\Bigr) = \|\boldsymbol{\xi}\|^2 \max_{\rho:\ \|\rho\boldsymbol{\xi}\|\le \mathtt{r}} \Bigl(\frac{\tau\|\boldsymbol{\xi}\|}{3}\rho^3 - (\rho-1)^2\Bigr).$$

Define $a = \tau\|\boldsymbol{\xi}\|$. The conditions $\|\boldsymbol{\xi}\| \le 2\mathtt{r}/3$ and $\tau\,\mathtt{r} \le 1/2$ imply $a \le 1/3$ and $\|\rho\boldsymbol{\xi}\| \le \mathtt{r}$ implies $|\rho| \le 3/2$. The function $a\rho^3/3 - (\rho-1)^2$ is concave on the interval $|\rho| \le 3/2$ and hence, the maximizer ρ fulfills $a\rho^2 - 2\rho + 2 = 0$ yielding

$$\rho = \frac{1 \pm \sqrt{1-2a}}{a}\,, \qquad |\rho| \le 3/2.$$

As $a \in [0, 1/3]$, we can only use

$$\rho_a = \frac{1 - \sqrt{1-2a}}{a} = \frac{2}{1+\sqrt{1-2a}}, \quad \rho_a - 1 = \frac{2a}{(1+\sqrt{1-2a})^2}.$$

Therefore,

$$\max_{\|\boldsymbol{u}\| \le \mathtt{r}} \left(\frac{\tau}{3} \|\boldsymbol{u}\|^3 - \|\boldsymbol{u} - \boldsymbol{\xi}\|^2 \right) = \frac{\tau \|\boldsymbol{\xi}\|^3 \rho_a^3}{3} - \|\boldsymbol{\xi}\|^2 (\rho_a - 1)^2$$

$$= \frac{\tau \|\boldsymbol{\xi}\|^3}{3} \frac{8(1+\sqrt{1-2a}) - 12a}{(1+\sqrt{1-2a})^4} \le \frac{\tau \|\boldsymbol{\xi}\|^3}{3} \max_{a \in [0,1/3]} \frac{8(1+\sqrt{1-2a}) - 12a}{(1+\sqrt{1-2a})^4} \le \frac{\tau \|\boldsymbol{\xi}\|^3}{2}$$

With $y = 1 + \sqrt{1-2a}$ or $-2a = (y-1)^2 - 1 = y^2 - 2y$, represent

$$\phi(a) \stackrel{\text{def}}{=} \frac{8(1+\sqrt{1-2a}) - 12a}{(1+\sqrt{1-2a})^4} = \frac{8y + 6y^2 - 12y}{y^4} = \frac{6y - 4}{y^3},$$

and the latter decreases with $y \ge 1$. As $\phi(1/3) \le 3/2$, (19) follows.

The proof of (20) is similar. The general case can be reduced to the univariate one by using $\boldsymbol{u} = \rho\boldsymbol{\xi}$. With $a = \tau\|\boldsymbol{\xi}\|$, the minimizer ρ_a reads as

$$\rho_a = \frac{2}{1+\sqrt{1+2a}}, \quad 1-\rho_a = \frac{\sqrt{1+2a}-1}{\sqrt{1+2a}+1} = \frac{2a}{(\sqrt{1+2a}+1)^2},$$

yielding for $a \in [0, 1/3]$

$$\min_{\|\boldsymbol{u}\|\le \mathbf{r}} \left(\frac{\tau}{3}\|\boldsymbol{u}\|^3 + \|\boldsymbol{u}-\boldsymbol{\xi}\|^2\right) = \frac{\tau\|\boldsymbol{\xi}\|^3\rho_a^3}{3} + \|\boldsymbol{\xi}\|^2(\rho_a-1)^2$$

$$\le \frac{\tau\|\boldsymbol{\xi}\|^3}{3} \max_{a\in[0,1/3]} \frac{8(1+\sqrt{1+2a})+12a}{(1+\sqrt{1+2a})^4},$$

and with $y = 1+\sqrt{1+2a}$ or $2a = y^2-2y$,

$$\max_{a\in[0,1/3]} \frac{8(1+\sqrt{1+2a})+12a}{(1+\sqrt{1+2a})^4} \le \max_{y\ge 2} \frac{8y+6y^2-12y}{y^4} = \max_{y\ge 2}\frac{6y-4}{y^3} = 1.$$

Proposition

Let $f(\boldsymbol{\upsilon})$ *be a strongly concave function with* $f(\boldsymbol{\upsilon}^*) = \max_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon})$ *and* $\mathbb{F} = -\nabla^2 f(\boldsymbol{\upsilon}^*)$ *. Assume* $(\boldsymbol{\mathcal{T}_3^*})$ *at* $\boldsymbol{\upsilon}^*$ *with* $\mathbb{D}^2$ *,* $\mathtt{r}$ *, and* τ_3 *such that*

$$\mathbb{D}^2 \le \mathbb{F}, \quad \mathtt{r} \ge \frac{3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|, \quad \tau_3\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| < \frac{4}{9}.$$

Then $\|\mathbb{D}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\| \le (3/2)\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|$ *and moreover,*

$$\|\mathbb{D}^{-1}\mathbb{F}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^* - \mathbb{F}^{-1}\boldsymbol{A})\| \le \frac{3\tau_3}{4}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2. \tag{21}$$

If the function f is quadratic and concave with the maximum at $\boldsymbol{\upsilon}^*$ then the linearly perturbed function g is also quadratic and concave with the maximum at $\breve{\boldsymbol{\upsilon}} = \boldsymbol{\upsilon}^* + \mathbb{F}^{-1}\boldsymbol{A}$.

In general, the point $\breve{\boldsymbol{\upsilon}}$ is not the maximizer of g, however, it is very close to $\boldsymbol{\upsilon}^\circ$. We use that $\nabla f(\boldsymbol{\upsilon}^*) = 0$ and $-\nabla^2 f(\boldsymbol{\upsilon}^*) = \mathbb{F}$. Then (27) of Lemma 4 yields

$$\begin{aligned} \bigl\| \mathbb{D}^{-1} \nabla g(\breve{\boldsymbol{\upsilon}}) \bigr\| &= \bigl\| \mathbb{D}^{-1} \{ \nabla f(\boldsymbol{\upsilon}^* + \mathbb{F}^{-1}\boldsymbol{A}) - \nabla f(\boldsymbol{\upsilon}^*) + \boldsymbol{A} \} \bigr\| \\ &\leq \frac{\tau_3}{2} \| \mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A} \|^2 . \end{aligned} \tag{22}$$

As $\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \le 2\mathtt{r}/3$, condition $(\boldsymbol{\mathcal{T}_3^*})$ can be applied in the $\mathtt{r}/3$-vicinity of $\breve{\boldsymbol{\upsilon}}$. Fix any $\boldsymbol{\upsilon}$ with $\|\mathbb{D}(\boldsymbol{\upsilon}-\breve{\boldsymbol{\upsilon}})\| \le \mathtt{r}/3$ and define $\Delta = \boldsymbol{\upsilon}-\breve{\boldsymbol{\upsilon}}$. By (29) of Lemma 4

$$\bigl\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{\upsilon}) - \nabla g(\breve{\boldsymbol{\upsilon}}) + \mathbb{F}\Delta\}\bigr\| = \bigl\|\mathbb{D}^{-1}\{\nabla f(\boldsymbol{\upsilon}) - \nabla f(\breve{\boldsymbol{\upsilon}}) + \mathbb{F}\Delta\}\bigr\|$$

$$\le \frac{3\tau_3}{2}\|\mathbb{D}\Delta\|^2 .$$

In particular, this and (22) yield

$$\bigl\|\mathbb{D}^{-1}\{\nabla g(\breve{\boldsymbol{\upsilon}}+\Delta) + \mathbb{F}\Delta\}\bigr\| \le 2\tau_3\|\mathbb{D}\Delta\|^2 .$$

For any $\boldsymbol{u}$ with $\|\boldsymbol{u}\| = 1$, this implies

$$\bigl|\bigl\langle \nabla g(\breve{\boldsymbol{\upsilon}}+\Delta) + \mathbb{F}\Delta, \mathbb{D}^{-1}\boldsymbol{u}\bigr\rangle\bigr| \le 2\tau_3\|\mathbb{D}\Delta\|^2 . \qquad (23)$$

Suppose now that $\|\mathbb{D}\varDelta\| = \mathtt{r}/3$ and consider the function $h(t) = g(\breve{\boldsymbol{\upsilon}} + t\varDelta)$. Then $h'(t) = \langle \nabla g(\breve{\boldsymbol{\upsilon}} + t\varDelta), \varDelta \rangle$ and (23) implies with $\boldsymbol{u} = \mathbb{D}\varDelta / \|\mathbb{D}\varDelta\|$

$$\bigl| \langle \nabla g(\breve{\boldsymbol{\upsilon}} + \varDelta), \varDelta \rangle + \|\mathbb{F}^{1/2}\varDelta\|^2 \bigr| \leq 2\tau_3 \|\mathbb{D}\varDelta\|^3 .$$

As $\mathbb{F} \geq \mathbb{D}^2$, this yields

$$h'(1) \leq 2\tau_3 \|\mathbb{D}\varDelta\|^3 - \|\mathbb{D}\varDelta\|^2. \tag{24}$$

Similarly, (22) yields by $\|\mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A}\| = 2\mathtt{r}/3$

$$|h'(0)| = \bigl| \langle \nabla g(\breve{\boldsymbol{\upsilon}}), \varDelta \rangle \bigr| \leq \frac{\tau_3}{2} \|\mathbb{D}\, \mathbb{F}^{-1} \boldsymbol{A}\|^2 \, \|\mathbb{D}\varDelta\| = \frac{2\tau_3}{9} \, \mathtt{r}^2 \, \|\mathbb{D}\varDelta\| . \tag{25}$$

Concavity of $g(\cdot)$ ensures that $t^* = \operatorname{argmax}_t h(t)$ satisfies $|t^*| \leq 1$ if

$$h'(1) < -|h'(0)|, \qquad h'(-1) < |h'(0)|.$$

Due to (24), (25), and $\|\mathbb{D}\varDelta\| = \mathtt{r}/3$, the latter condition reads

$$\frac{2\tau_3}{9}\mathtt{r}^2 \, \|\mathbb{D}\varDelta\| + 2\tau_3\|\mathbb{D}\varDelta\|^3 - \|\mathbb{D}\varDelta\|^2 = \|\mathbb{D}\varDelta\| \, \mathtt{r}\Big(\frac{2\tau_3 \, \mathtt{r}}{9} + \frac{2\tau_3 \, \mathtt{r}}{9} - \frac{1}{3}\Big) < 0.$$

which is fulfilled because of $\tau_3\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \le 4/9$ and $\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| = 2\mathtt{r}/3$. We summarize that $\boldsymbol{\upsilon}^{\circ} = \operatorname{argmax}_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon})$ satisfies $\|\mathbb{D}\,(\boldsymbol{\upsilon}^{\circ} - \breve{\boldsymbol{\upsilon}})\| \le \mathtt{r}/3$ while $\|\mathbb{D}(\breve{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^{*})\| = \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| = 2\mathtt{r}/3$. Therefore,

$$\|\mathbb{D}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\| \le \mathtt{r}.$$

This allows us to use $(\boldsymbol{\mathcal{T}_3^*})$ at this point for establishing (21). By definition $\nabla g(\boldsymbol{\upsilon}^{\circ}) = 0$ and hence,

$$\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{\upsilon}^{*} + \mathbb{F}^{-1}\boldsymbol{A}) - \nabla g(\boldsymbol{\upsilon}^{\circ})\}\| \le \frac{\tau_3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 . \qquad (26)$$

By (29) of Lemma 4, it holds with $\varDelta = \boldsymbol{\upsilon}^{*} + \mathbb{F}^{-1}\boldsymbol{A} - \boldsymbol{\upsilon}^{\circ}$

$$\left\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{\upsilon}^{*} + \mathbb{F}^{-1}\boldsymbol{A}) - \nabla g(\boldsymbol{\upsilon}^{\circ}) - \nabla^{2} g(\boldsymbol{\upsilon}^{*})\varDelta\}\right\| \leq \frac{3\tau_3}{2}\|\mathbb{D}\varDelta\|^{2} .$$

Combining with (26) yields

$$\|\mathbb{D}^{-1}\mathbb{F}\varDelta\| \leq \frac{3\tau_3}{2}\|\mathbb{D}\varDelta\|^{2} + \frac{\tau_3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^{2} \leq \frac{3\tau_3}{2}\|\mathbb{D}^{-1}\mathbb{F}\varDelta\|^{2} + \frac{\tau_3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^{2} .$$

As $2x \leq \alpha x^{2} + \beta$ with $\alpha = 3\tau_3$, $\beta = \tau_3\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^{2}$, and
$x = \|\mathbb{D}^{-1}\mathbb{F}\varDelta\| \in (0, 1/\alpha)$ implies $x \leq \beta/(2 - \alpha\beta)$, this yields

$$\|\mathbb{D}^{-1}\mathbb{F}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*} - \mathbb{F}^{-1}\boldsymbol{A})\| \leq \frac{\tau_3}{2 - 3\tau_3^{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^{2}}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^{2}$$

and (21) follows by $\tau_3\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \leq 4/9$.

Lemma

Assume $(\mathcal{T}_3^*)$ *at* $\boldsymbol{\upsilon}$ *. Let* $\mathcal{U}_{\mathtt{r}} = \{\boldsymbol{u} \colon \|\mathbb{D}\boldsymbol{u}\| \leq \mathtt{r}\}$ *. Then*

$$\left\|\mathbb{D}^{-1}\left\{\nabla f(\boldsymbol{\upsilon}+\boldsymbol{u}) - \nabla f(\boldsymbol{\upsilon}) - \langle \nabla^2 f(\boldsymbol{\upsilon}), \boldsymbol{u}\rangle\right\}\right\| \leq \frac{\tau_3}{2} \|\mathbb{D}\boldsymbol{u}\|^2 , \quad \boldsymbol{u} \in \mathcal{U}_{\mathtt{r}} . \tag{27}$$

Also for all $\boldsymbol{u}, \boldsymbol{u}_1 \in \mathcal{U}_{\mathtt{r}}$

$$\left\|\mathbb{D}^{-1}\left\{\nabla^2 f(\boldsymbol{\upsilon}+\boldsymbol{u}_1) - \nabla^2 f(\boldsymbol{\upsilon}+\boldsymbol{u})\right\}\mathbb{D}^{-1}\right\| \leq \tau_3 \|\mathbb{D}(\boldsymbol{u}_1 - \boldsymbol{u})\| \quad (28$$

$$\left\|\mathbb{D}^{-1}\left\{\nabla f(\boldsymbol{\upsilon}+\boldsymbol{u}_1) - \nabla f(\boldsymbol{\upsilon}+\boldsymbol{u}) - \nabla^2 f(\boldsymbol{\upsilon})(\boldsymbol{u}_1 - \boldsymbol{u})\right\}\right\| \leq \frac{3\tau_3}{2} \|\mathbb{D}(\boldsymbol{u}_1 - \boldsymbol{u})\|^2 . \quad (29$$

Moreover, under $(\mathcal{T}_4^*)$ *, for any* $\boldsymbol{u} \in \mathcal{U}_{\mathtt{r}}$ *,*

$$\left\|\mathbb{D}^{-1}\left\{\nabla f(\boldsymbol{\upsilon}+\boldsymbol{u}) - \nabla f(\boldsymbol{\upsilon}) - \langle \nabla^2 f(\boldsymbol{\upsilon}), \boldsymbol{u}\rangle - \frac{1}{2}\langle \nabla^3 f(\boldsymbol{\upsilon}), \boldsymbol{u}^{\otimes 2}\rangle\right\}\right\| \leq \frac{\tau_4}{6} \|\mathbb{D}\boldsymbol{u}\|^3 . \ ($$

Denote

$$\boldsymbol{A} \stackrel{\text{def}}{=} \nabla f(\boldsymbol{v}+\boldsymbol{u}) - \nabla f(\boldsymbol{v}) - \langle \nabla^2 f(\boldsymbol{v}), \boldsymbol{u} \rangle .$$

For any vector $\boldsymbol{w} \in \mathbb{R}^p$, $(\mathcal{T}_3^*)$ and (12) imply

$$\bigl|\langle \boldsymbol{A}, \boldsymbol{w} \rangle\bigr| \leq \frac{\tau_3}{2} \|\mathbb{D}\boldsymbol{u}\|^2 \|\mathbb{D}\boldsymbol{w}\|.$$

Therefore,

$$\|\mathbb{D}^{-1}\boldsymbol{A}\| = \sup_{\|\boldsymbol{w}\|=1} \bigl|\langle \mathbb{D}^{-1}\boldsymbol{A}, \boldsymbol{w} \rangle\bigr| = \sup_{\|\boldsymbol{w}\|=1} \bigl|\langle \boldsymbol{A}, \mathbb{D}^{-1}\boldsymbol{w} \rangle\bigr| \leq \frac{\tau_3}{2} \|\mathbb{D}\boldsymbol{u}\|^2$$

which yields the first statement.

For (30), apply

$$\boldsymbol{A} \stackrel{\text{def}}{=} \nabla f(\boldsymbol{v}+\boldsymbol{u}) - \nabla f(\boldsymbol{v}) - \langle \nabla^2 f(\boldsymbol{v}), \boldsymbol{u} \rangle - \frac{1}{2}\langle \nabla^3 f(\boldsymbol{v}), \boldsymbol{u}^{\otimes 2} \rangle$$

and use $(\mathcal{T}_4^*)$ and (13) instead of $(\mathcal{T}_3^*)$ and (12).

Further, with $\mathbb{B}_1 \stackrel{\mathrm{def}}{=} \nabla^2 f(\boldsymbol{\upsilon}+\boldsymbol{u}_1) - \nabla^2 f(\boldsymbol{\upsilon}+\boldsymbol{u})$ and $\Delta = \boldsymbol{u}_1 - \boldsymbol{u}$, by $(\boldsymbol{\mathcal{T}_3^*})$, for any $\boldsymbol{w} \in \mathbb{R}^p$ and some $t \in [0,1]$,

$$\bigl|\langle \mathbb{D}^{-1}\bigl\{\nabla^2 f(\boldsymbol{\upsilon}+\boldsymbol{u}_1) - \nabla^2 f(\boldsymbol{\upsilon}+\boldsymbol{u})\bigr\}\,\mathbb{D}^{-1}, \boldsymbol{w}^{\otimes 2}\rangle\bigr| = \bigl|\langle \mathbb{B}_1, (\mathbb{D}^{-1}\boldsymbol{w})^{\otimes 2}\rangle\bigr|$$
$$= \bigl|\bigl\langle \nabla^3 f(\boldsymbol{\upsilon}+\boldsymbol{u}+t\Delta), \Delta \otimes (\mathbb{D}^{-1}\boldsymbol{w})^{\otimes 2}\bigr\rangle\bigr| \le \tau_3 \|\mathbb{D}\Delta\|\,\|\boldsymbol{w}\|^2 .$$

This proves (28). Similarly, for some $t \in [0,1]$

$$\bigl|\bigl\langle \mathbb{D}^{-1}\bigl\{\nabla f(\boldsymbol{\upsilon}+\boldsymbol{u}_1) - \nabla f(\boldsymbol{\upsilon}+\boldsymbol{u})\bigr\} - \nabla^2 f(\boldsymbol{\upsilon}+\boldsymbol{u})\Delta\bigr\}, \boldsymbol{w}\bigr\rangle\bigr|$$
$$= \frac{1}{2}\bigl|\bigl\langle \nabla^3 f(\boldsymbol{\upsilon}+\boldsymbol{u}+t\Delta), \Delta \otimes \Delta \otimes \mathbb{D}^{-1}\boldsymbol{w}\bigr\rangle\bigr| \le \frac{\tau_3}{2}\|\mathbb{D}\Delta\|^2\,\|\boldsymbol{w}\|$$

and with $\mathbb{B} = \nabla^2 f(\boldsymbol{\upsilon}+\boldsymbol{u}) - \nabla^2 f(\boldsymbol{\upsilon})$, by (28),

$$\bigl\|\mathbb{D}^{-1}\mathbb{B}\Delta\bigr\| \le \|\mathbb{D}^{-1}\mathbb{B}\,\mathbb{D}^{-1}\|\;\|\mathbb{D}\Delta\| \le \tau_3\|\mathbb{D}\Delta\|^2 .$$

This completes the proof of (29).

Lemma

If $2x \leq \alpha x^2 + \beta$ *and* $x \in (0, 1/\alpha)$ *for* $\alpha\beta \leq 1$ *, then*

$$x \leq \frac{\beta}{2 - \alpha\beta} \, .$$

The roots of $\alpha x^2 + \beta = 2x$ satisfy

$$x = \frac{1 \pm \sqrt{1 - \alpha\beta}}{\alpha}$$

As $x \leq 1/\alpha$, we only consider

$$x \leq \frac{1 - \sqrt{1 - \alpha\beta}}{\alpha} = \frac{\alpha\beta}{\alpha(1 + \sqrt{1 - \alpha\beta})} \leq \frac{\beta}{1 + 1 - \alpha\beta}\,.$$

Lemma

Assume $\mathbb{D}^2 \le \mathbb{F}$ *and let some other matrix* $\mathbb{F}_1 \in \mathfrak{M}_p$ *satisfy*

$$\|\mathbb{D}^{-1}\,(\mathbb{F}_1 - \mathbb{F})\,\mathbb{D}^{-1}\| \le \omega \tag{31}$$

with $\omega < 1$ *. Then for any vector* $\boldsymbol{u}$

$$\|\mathbb{F}^{-1/2}\,(\mathbb{F}_1 - \mathbb{F})\,\mathbb{F}^{-1/2}\| \le \omega\,, \tag{32}$$

$$\|\mathbb{F}^{1/2}\,(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\,\mathbb{F}^{1/2}\| \le \frac{\omega}{1-\omega}\,, \tag{33}$$

$$\frac{1}{1+\omega}\,\|\mathbb{D}\,\mathbb{F}^{-1}\,\mathbb{D}\| \le \|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{D}\| \le \frac{1}{1-\omega}\,\|\mathbb{D}\,\mathbb{F}^{-1}\,\mathbb{D}\|\,, \tag{34}$$

$$(1-\omega)\|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\| \le \|\mathbb{D}^{-1}\mathbb{F}_1\boldsymbol{u}\| \le (1+\omega)\|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\|\,, \tag{35}$$

$$\frac{1-2\omega}{1-\omega}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{u}\| \le \|\mathbb{D}\,\mathbb{F}_1^{-1}\boldsymbol{u}\| \le \frac{1}{1-\omega}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{u}\|\,. \tag{36}$$

Statement (32) follows from (31) because of $\mathbb{F}^{-1} \le \mathbb{D}^{-2}$. Define now $\boldsymbol{U} \stackrel{\text{def}}{=} \mathbb{F}^{-1/2}\,(\mathbb{F}_1 - \mathbb{F})\,\mathbb{F}^{-1/2}$. Then $\|\boldsymbol{U}\| \le \omega$ and

$$\|\mathbb{F}^{1/2}\,(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\,\mathbb{F}^{1/2}\| = \|(I\!\!I + \boldsymbol{U})^{-1} - I\!\!I\| \le \frac{1}{1-\omega}\|\boldsymbol{U}\|$$

yielding (33). Further,

$$\begin{aligned}
\|\mathbb{D}\,(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\,\mathbb{D}\| &= \|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{F}_1(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\,\mathbb{F}\,\mathbb{F}^{-1}\mathbb{D}\| \\
&= \|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{D}\,\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\,\mathbb{D}\,\mathbb{F}^{-1}\mathbb{D}\| \\
&\le \|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{D}\|\,\|\mathbb{D}\,\mathbb{F}^{-1}\mathbb{D}\|\,\|\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\| \le \omega\|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{D}\|\,.
\end{aligned}$$

This implies (34).

Also, by $\mathbb{D}^2 \leq \mathbb{F}$

$$\|\mathbb{D}^{-1}\mathbb{F}_1\boldsymbol{u}\| \leq \|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\| + \|\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\mathbb{D}\boldsymbol{u}\| \leq \|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\| + \omega\,\|\mathbb{D}\boldsymbol{u}\|$$
$$\leq \|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\| + \omega\|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\| \leq (1+\omega)\|\mathbb{D}^{-1}\mathbb{F}\boldsymbol{u}\|,$$

and (35) follows. Similarly

$$\begin{aligned}
\|\mathbb{D}\,(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\boldsymbol{u}\| &= \|\mathbb{D}\,\mathbb{F}_1^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{F}^{-1}\boldsymbol{u}\| \\
&= \|\mathbb{D}\,\mathbb{F}_1^{-1}\mathbb{D}\;\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\,\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{u}\| \\
&\leq \|\mathbb{D}^{-1}\,(\mathbb{F}_1 - \mathbb{F})\,\mathbb{D}^{-1}\|\;\|\mathbb{D}\,\mathbb{F}_1^{-1}\,\mathbb{D}\|\;\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{u}\| \\
&\leq \frac{\omega}{1-\omega}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{u}\|
\end{aligned}$$

and (36) follows as well.

Proposition

Let $f(\boldsymbol{\upsilon})$ *be a strongly concave function with* $f(\boldsymbol{\upsilon}^*) = \max_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon})$ *and* $\mathbb{F} = -\nabla^2 f(\boldsymbol{\upsilon}^*)$ *, and let* $f(\boldsymbol{\upsilon})$ *follow* $(\boldsymbol{\mathcal{T}_3^*})$ *and* $(\boldsymbol{\mathcal{T}_4^*})$ *with some* $\mathbb{D}^2$ *,* τ_3 *,* τ_4 *, and* $\mathtt{r}$ *satisfying*

$$\mathbb{D}^2 \le \mathbb{F}, \quad \mathtt{r} = \frac{3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|, \quad \tau_3\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| < \frac{4}{9}, \quad \tau_4\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 < \frac{1}{3}. \quad (37)$$

Let $g(\boldsymbol{\upsilon})$ *fulfill* (1) *with some vector* $\boldsymbol{A}$ *and* $g(\boldsymbol{\upsilon}^\circ) = \max_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon})$ *. Then* $\|\mathbb{D}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*)\| \le (3/2)\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|$ *. Further, define*

$$\boldsymbol{a} = \mathbb{F}^{-1}\{\boldsymbol{A} + \nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\}, \quad (38)$$

where $\mathcal{T}(\boldsymbol{u}) = \frac{1}{6}\langle\nabla^3 f(\boldsymbol{\upsilon}^*), \boldsymbol{u}^{\otimes 3}\rangle$ *for* $\boldsymbol{u} \in \mathbb{R}^p$ *. Then*

$$\|\mathbb{D}^{-1}\mathbb{F}(\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^* - \boldsymbol{a})\| \le (\tau_4/2 + \tau_3^2)\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3. \quad (39)$$

Proposition 5 yields (21). By $(\mathcal{T}_3^*)$

$$\|\mathbb{D}^{-1}\,\mathbb{F}(\boldsymbol{a}-\mathbb{F}^{-1}\boldsymbol{A})\| = \|\mathbb{D}^{-1}\,\nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\|$$

$$= \sup_{\|\boldsymbol{u}\|=1} 3\big|\langle\mathcal{T},\mathbb{F}^{-1}\boldsymbol{A}\otimes\mathbb{F}^{-1}\boldsymbol{A}\otimes\mathbb{D}^{-1}\boldsymbol{u}\rangle\big| \le \frac{\tau_3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 . \tag{40}$$

As $\mathbb{D}^{-1}\,\mathbb{F}\ge\mathbb{F}^{1/2}\ge\mathbb{D}$, this implies by $\tau_3\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|\le 4/9$

$$\|\mathbb{D}\boldsymbol{a}\| \le \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| + \|\mathbb{D}\,\mathbb{F}^{-1}\,\nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\|$$

$$\le \Big(1+\frac{\tau_3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|\Big)\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \le \frac{11}{9}\,\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\| \tag{41}$$

and

$$\|\mathbb{F}^{1/2}\boldsymbol{a}-\mathbb{F}^{-1/2}\boldsymbol{A}\| \le \frac{\tau_3}{2}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 .$$

Next, again by $(\mathcal{T}_3^*)$, for any $\boldsymbol{w}$

$$\|\mathbb{D}^{-1}\, \nabla^2 \mathcal{T}(\boldsymbol{w})\, \mathbb{D}^{-1}\| \;=\; \sup_{\|\boldsymbol{u}\|=1} 6 \bigl| \langle \mathcal{T}, \boldsymbol{w} \otimes (\mathbb{D}^{-1} \boldsymbol{u})^{\otimes 2} \rangle \bigr| \le \tau_3 \|\mathbb{D} \boldsymbol{w}\| \,.$$

The tensor $\nabla^2 \mathcal{T}(\boldsymbol{u})$ is linear in $\boldsymbol{u}$, hence for any $t \in [0,1]$

$$\begin{aligned}
&\|\mathbb{D}^{-1}\, \nabla^2 \mathcal{T}(t\boldsymbol{a} + (1-t)\mathbb{F}^{-1}\boldsymbol{A})\, \mathbb{D}^{-1}\| \\
&\quad \le \max\{\|\mathbb{D}^{-1}\, \nabla^2 \mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\, \mathbb{D}^{-1}\|, \|\mathbb{D}^{-1} \nabla^2 \mathcal{T}(\boldsymbol{a}) \mathbb{D}^{-1}\|\} \\
&\quad \le \tau_3 \max\{\|\mathbb{D}\, \mathbb{F}^{-1}\boldsymbol{A}\|, \|\mathbb{D}\boldsymbol{a}\|\} \,.
\end{aligned}$$

Based on (41), assume $\|\mathbb{D}\, \mathbb{F}^{-1}\boldsymbol{A}\| \le \|\mathbb{D}\boldsymbol{a}\| \le (11/9) \|\mathbb{D}\, \mathbb{F}^{-1}\boldsymbol{A}\|$. Then (40) yield

$$\begin{aligned}
&\|\mathbb{D}^{-1} \nabla \mathcal{T}(\boldsymbol{a}) - \mathbb{D}^{-1} \nabla \mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\| \\
&\quad = \mathbb{D}^{-1}\, \nabla^2 \mathcal{T}(t\boldsymbol{a} + (1-t)\mathbb{F}^{-1}\boldsymbol{A})\, \mathbb{D}^{-1}\| \; \|\mathbb{D}\, \mathbb{F}^{-1}(\boldsymbol{a} - \mathbb{F}^{-1}\boldsymbol{A})\| \\
&\quad \le \frac{\tau_3^2}{2} \|\mathbb{D}\, \mathbb{F}^{-1}\boldsymbol{A}\|^2 \, \|\mathbb{D}\boldsymbol{a}\| \le \frac{2\tau_3^2}{3} \|\mathbb{D}\, \mathbb{F}^{-1}\boldsymbol{A}\|^3 \,.
\end{aligned}$$

Proof

Further, $-\nabla^2 f(0) = \mathbb{F}$, $\nabla\mathcal{T}(\boldsymbol{a}) = \frac{1}{2}\langle\nabla^3 f(0), \boldsymbol{a}\otimes\boldsymbol{a}\rangle$. By (30) and (41)

$$\bigl\|\mathbb{D}^{-1}\{\nabla f(\boldsymbol{a}) + \mathbb{F}\boldsymbol{a} - \nabla\mathcal{T}(\boldsymbol{a})\}\bigr\| \le \frac{\tau_4}{6}\|\mathbb{D}\boldsymbol{a}\|^3 \le \frac{(11/9)^3\tau_4}{6}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3$$
$$\le \frac{\tau_4}{3}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3 .$$

Next we bound $\bigl\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{a}) - \nabla g(\boldsymbol{v}^\circ)\}\bigr\|$. As $\nabla g(\boldsymbol{v}^\circ) = 0$, (1) and (38) imply

$$\bigl\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{a}) - \nabla g(\boldsymbol{v}^\circ)\}\bigr\| = \bigl\|\mathbb{D}^{-1}\nabla g(\boldsymbol{a})\bigr\| = \bigl\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{a}) + \mathbb{F}\boldsymbol{a} - \nabla\mathcal{T}(\boldsymbol{A}) - \boldsymbol{A}\}\bigr\|$$
$$\le \bigl\|\mathbb{D}^{-1}\{\nabla f(\boldsymbol{a}) + \mathbb{F}\boldsymbol{a} - \nabla\mathcal{T}(\boldsymbol{a})\}\bigr\| + \|\mathbb{D}^{-1}\{\nabla\mathcal{T}(\boldsymbol{a}) - \nabla\mathcal{T}(\boldsymbol{A})\}\| \le \diamondsuit_1 , \tag{42}$$

where $\diamondsuit_1 \stackrel{\text{def}}{=} \frac{\tau_4 + 2\tau_3^2}{3}\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3$, and by (37)

$$3\tau_3\,\diamondsuit_1 = \tau_3\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|\;\tau_4\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^2 + 2\tau_3^3\|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3 < \frac{1}{3} . \tag{43}$$

Further, $\nabla^2 g(0) = \nabla^2 f(0) = -\mathbb{F}$, and (29) of Lemma 4 implies

$$\begin{aligned}
&\bigl\|\mathbb{D}^{-1}\{\nabla g(\boldsymbol{a}) - \nabla g(\boldsymbol{\upsilon}^{\circ}) + \mathbb{F}(\boldsymbol{a} - \boldsymbol{\upsilon}^{\circ})\}\bigr\| \\
&\qquad = \bigl\|\mathbb{D}^{-1}\{\nabla f(\boldsymbol{a}) - \nabla f(\boldsymbol{\upsilon}^{\circ}) + \mathbb{F}(\boldsymbol{a} - \boldsymbol{\upsilon}^{\circ})\}\bigr\| \le \frac{3\tau_3}{2} \|\mathbb{D}(\boldsymbol{a} - \boldsymbol{\upsilon}^{\circ})\|^2 .
\end{aligned}$$

Combining with (42) yields in view of $\mathbb{D}^2 \le \mathbb{F}$

$$\|\mathbb{D}^{-1}\mathbb{F}(\boldsymbol{a} - \boldsymbol{\upsilon}^{\circ})\| \le \frac{3\tau_3}{2} \|\mathbb{D}(\boldsymbol{a} - \boldsymbol{\upsilon}^{\circ})\|^2 + \diamondsuit_1 \le \frac{3\tau_3}{2} \|\mathbb{D}^{-1}\mathbb{F}(\boldsymbol{a} - \boldsymbol{\upsilon}^{\circ})\|^2 + \diamondsuit_1 .$$

As $2x \le \alpha x^2 + \beta$ with $\alpha = 3\tau_3$, $\beta = 2\diamondsuit_1$, and $x \in (0, 1/\alpha)$ implies $x \le \beta/(2 - \alpha\beta)$, we conclude by (43)

$$\|\mathbb{D}^{-1}\mathbb{F}(\boldsymbol{a} - \boldsymbol{\upsilon}^{\circ})\| \le \frac{\diamondsuit_1}{1 - 3\tau_3 \diamondsuit_1} \le \frac{\tau_4 + 2\tau_3^2}{2} \|\mathbb{D}\,\mathbb{F}^{-1}\boldsymbol{A}\|^3 ,$$

and (39) follows.

Here we discuss the case when $g(\boldsymbol{v}) - f(\boldsymbol{v})$ is quadratic.

The general case can be reduced to the situation with $g(\boldsymbol{v}) = f(\boldsymbol{v}) - \|G\boldsymbol{v}\|^2/2$. To make the dependence of G more explicit, denote $f_G(\boldsymbol{v}) \stackrel{\text{def}}{=} f(\boldsymbol{v}) - \|G\boldsymbol{v}\|^2/2$,

$$\boldsymbol{v}^* = \operatorname*{argmax}_{\boldsymbol{v}} f(\boldsymbol{v}),$$

$$\boldsymbol{v}_G^* = \operatorname*{argmax}_{\boldsymbol{v}} f_G(\boldsymbol{v}) = \operatorname*{argmax}_{\boldsymbol{v}} \bigl\{ f(\boldsymbol{v}) - \|G\boldsymbol{v}\|^2/2 \bigr\}.$$

We study the bias $\boldsymbol{v}_G^* - \boldsymbol{v}^*$ induced by this penalization.

Lemma

Let $f(\boldsymbol{\upsilon})$ *be quadratic with* $\mathbb{F} \equiv -\nabla^2 f(\boldsymbol{\upsilon})$ *. Define*

$$\boldsymbol{S}_G \equiv -G^2 \boldsymbol{\upsilon}^*.$$

Then it holds with $\mathbb{F}_G = \mathbb{F} + G^2$

$$\begin{aligned} \boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^* &= \mathbb{F}_G^{-1} \boldsymbol{S}_G = -\mathbb{F}_G^{-1} G^2 \boldsymbol{\upsilon}^*, \\ f_G(\boldsymbol{\upsilon}_G^*) - f_G(\boldsymbol{\upsilon}^*) &= \frac{1}{2} \|\mathbb{F}_G^{-1/2} \boldsymbol{S}_G\|^2 = \frac{1}{2} \|\mathbb{F}_G^{-1/2} G^2 \boldsymbol{\upsilon}^*\|^2 . \end{aligned}$$

Quadraticity of $f(\boldsymbol{\upsilon})$ implies quadraticity of $f_G(\boldsymbol{\upsilon})$ with $\nabla^2 f_G(\boldsymbol{\upsilon}) \equiv -\mathbb{F}_G$ and

$$\nabla f_G(\boldsymbol{\upsilon}^*) - \nabla f_G(\boldsymbol{\upsilon}_G^*) = \mathbb{F}_G\,(\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*).$$

Further, $\nabla f(\boldsymbol{\upsilon}^*) = 0$ yielding $\nabla f_G(\boldsymbol{\upsilon}^*) = \boldsymbol{S}_G = -G^2\boldsymbol{\upsilon}^*$. Together with $\nabla f_G(\boldsymbol{\upsilon}_G^*) = 0$, this implies

$$\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^* = \mathbb{F}_G^{-1}\boldsymbol{S}_G.$$

The Taylor expansion of f_G at $\boldsymbol{\upsilon}_G^*$ yields

$$f_G(\boldsymbol{\upsilon}^*) - f_G(\boldsymbol{\upsilon}_G^*) = -\frac{1}{2}\|\mathbb{F}_G^{1/2}(\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^*)\|^2 = -\frac{1}{2}\|\mathbb{F}_G^{-1/2}\boldsymbol{S}_G\|^2$$

and the assertion follows.

Proposition

Let $f_G(\boldsymbol{\upsilon}) = f(\boldsymbol{\upsilon}) - \|G\boldsymbol{\upsilon}\|^2/2$ *be concave and follow* $(\boldsymbol{\mathcal{T}_3^*})$ *with some* $\mathbb{D}^2$, τ_3, *and* $\mathtt{r}$ *satisfying*

$$\mathbb{D}^2 \le \mathbb{F}_G\,, \qquad \mathtt{r} \ge 3\mathsf{b}_G/2\,, \qquad \tau_3\,\mathsf{b}_G < 4/9,$$

where $\mathsf{b}_G = \|\mathbb{D}\,\mathbb{F}_G^{-1}\,G^2\boldsymbol{\upsilon}^*\|$. *Then*

$$\|\mathbb{D}(\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*)\| \le 3\mathsf{b}_G/2.$$

Moreover,

$$\bigl\|\mathbb{D}^{-1}\mathbb{F}_G(\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^* + \mathbb{F}_G^{-1}G^2\boldsymbol{\upsilon}^*)\bigr\| \le \frac{3\tau_3}{4}\,\mathsf{b}_G^2\,,$$

$$\Bigl|2f_G(\boldsymbol{\upsilon}_G^*) - 2f_G(\boldsymbol{\upsilon}^*) - \frac{1}{2}\|\mathbb{F}_G^{-1/2}G^2\boldsymbol{\upsilon}^*\|^2\Bigr| \le \frac{\tau_3}{2}\,\mathsf{b}_G^3\,.$$

Define $g_G(\boldsymbol{v})$ by

$$g_G(\boldsymbol{v}) - g_G(\boldsymbol{v}_G^*) = f_G(\boldsymbol{v}) - f_G(\boldsymbol{v}_G^*) + \langle G^2\boldsymbol{v}^*, \boldsymbol{v} - \boldsymbol{v}_G^* \rangle. \quad (44)$$

The function f_G is concave, the same holds for g_G from (44).

Hence, $\nabla g_G(\boldsymbol{v}^*) = 0$ implies $\boldsymbol{v}^* = \operatorname{argmax} g_G(\boldsymbol{v})$. By definition, $\nabla f(\boldsymbol{v}^*) = 0$ yielding $\nabla f_G(\boldsymbol{v}^*) = -G^2\boldsymbol{v}^* + G^2\boldsymbol{v}^* = 0$.

Now the results follow from Propositions 5 and 3 applied with $f(\boldsymbol{v}) = g_G(\boldsymbol{v}) = f_G(\boldsymbol{v}) - \langle \boldsymbol{A}, \boldsymbol{v} \rangle$, $g(\boldsymbol{v}) = f_G(\boldsymbol{v})$, and $\boldsymbol{A} = G^2\boldsymbol{v}^*$.

Define $\mathbb{F}_G = -\nabla^2 f(\boldsymbol{\upsilon}^*) + G^2$, $\boldsymbol{S}_G = G^2 \boldsymbol{\upsilon}^*$, and

$$\boldsymbol{m}_G = \mathbb{F}_G^{-1} \{ \boldsymbol{S}_G + \nabla \mathcal{T}(\mathbb{F}_G^{-1} \boldsymbol{S}_G) \}$$

with $\mathcal{T}(\boldsymbol{u}) = \frac{1}{6} \langle \nabla^3 f(\boldsymbol{\upsilon}^*), \boldsymbol{u}^{\otimes 3} \rangle$.

$(\mathcal{T}_4^*)$ $f(\boldsymbol{\upsilon})$ *is strongly concave,* $\mathbb{D}^2 \leq \nabla^2 f(\boldsymbol{\upsilon})$ *, and*

$$\sup_{\boldsymbol{u} \colon \|\mathbb{D}\boldsymbol{u}\| \leq \mathtt{r}} \sup_{\boldsymbol{z} \in \mathbb{R}^p} \frac{\bigl| \langle \nabla^4 f(\boldsymbol{\upsilon} + \boldsymbol{u}), \boldsymbol{z}^{\otimes 4} \rangle \bigr|}{\| \mathbb{D} \boldsymbol{z} \|^4} \leq \tau_4 .$$

Typically $\tau_3 \asymp n^{-1/2}$ and $\tau_4 \asymp n^{-1}$.

Proposition

Let f *be concave and* $\boldsymbol{\upsilon}^{*} = \operatorname{argmax}_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon})$ *. With* $\mathbb{F}_{G} = -\nabla^{2} f(\boldsymbol{\upsilon}^{*}) + G^{2}$ *. Let* $f(\boldsymbol{\upsilon})$ *follow* $(\boldsymbol{\mathcal{T}_{3}^{*}})$ *and* $(\boldsymbol{\mathcal{T}_{4}^{*}})$ *with some* $\mathbb{D}^{2}$ *,* τ_{3} *,* τ_{4} *, and* $\mathtt{r}$ *satisfying*

$$\mathbb{D}^{2} \le \mathbb{F}_{G} \, , \quad \mathtt{r} = \frac{3}{2} \mathsf{b}_{G} \, , \quad \tau_{3} \, \mathsf{b}_{G} < \frac{4}{9} \, , \quad \tau_{4} \, \mathsf{b}_{G}^{2} < \frac{1}{3} \, .$$

with $\mathsf{b}_{G} = \| \mathbb{D} \, \mathbb{F}_{G}^{-1} \, G^{2} \boldsymbol{\upsilon}^{*} \|$ *. Define*

$$\boldsymbol{m}_{G} = \mathbb{F}_{G}^{-1} \{ G^{2} \boldsymbol{\upsilon}^{*} + \nabla \mathcal{T} (\mathbb{F}_{G}^{-1} G^{2} \boldsymbol{\upsilon}^{*}) \}$$

with $\mathcal{T}(\boldsymbol{u}) = \frac{1}{6} \langle \nabla^{3} f(\boldsymbol{\upsilon}^{*}), \boldsymbol{u}^{\otimes 3} \rangle$ *and* $\nabla \mathcal{T} = \frac{1}{2} \langle \nabla^{3} f(\boldsymbol{\upsilon}^{*}), \boldsymbol{u}^{\otimes 2} \rangle$ *. Then*

$$\| \mathbb{D}^{-1} \mathbb{F}_{G} (\boldsymbol{\upsilon}^{*} - \boldsymbol{\upsilon}_{G}^{*} - \boldsymbol{m}_{G}) \| \le \frac{\tau_{4} + 2 \tau_{3}^{2}}{2} \, \mathsf{b}_{G}^{3} \, .$$

- Statistical inference for nonlinear regression. DNN training.
- Gaussian variational inference
- Bayesian optimization

Banach, S. (1938).
Über homogene Polynome in (L^2).
Studia Mathematica, 7(1):36–44.

Katsevich, A. and Rigollet, P. (2023).
On the approximation accuracy of gaussian variational inference.

Weierstraß-Institut für
Angewandte Analysis und Stochastik

Leibniz
Gemeinschaft

Linearly pertubed optimization: theory and applications

Vladimir Spokoiny ,
WIAS, HU Berlin

8. Oktober 2024

1. **Statistical inference**
   - Linear and SLS models
   - Nonlinear regression. Theoretical study
2. **Structural modeling: Examples**
   - Matrix completion
   - Gaussian mixture
   - Deep Neural Networks
3. **Gaussian Variational Inference**
4. **Optimization vs sampling**

Let $L(\boldsymbol{\upsilon})$ be a random function, $\boldsymbol{\upsilon} \in \Upsilon \subseteq \mathbb{R}^p$, $p < \infty$.
Given a quadratic penalty $\|G\boldsymbol{\upsilon}\|^2/2$, define

$$L_G(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \|G\boldsymbol{\upsilon}\|^2/2.$$

Consider

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmax}_{\boldsymbol{\upsilon}} L_G(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\upsilon}} \Bigl\{ L(\boldsymbol{\upsilon}) - \frac{1}{2}\|G\boldsymbol{\upsilon}\|^2 \Bigr\};$$

$$\boldsymbol{\upsilon}_G^* = \operatorname*{argmax}_{\boldsymbol{\upsilon}} \mathbb{E}\, L_G(\boldsymbol{\upsilon}) = \operatorname*{argmax}_{\boldsymbol{\upsilon}} \Bigl\{ \mathbb{E}\, L(\boldsymbol{\upsilon}) - \frac{1}{2}\|G\boldsymbol{\upsilon}\|^2 \Bigr\};$$

$$\boldsymbol{\upsilon}^* = \operatorname*{argmax}_{\boldsymbol{\upsilon}} \mathbb{E}\, L(\boldsymbol{\upsilon});$$

Aim: describe the estimation loss $\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*$ and the prediction loss (excess) $L_G(\widetilde{\boldsymbol{\upsilon}}_G) - L_G(\boldsymbol{\upsilon}^*)$.

- A linear model $\boldsymbol{Y} = \boldsymbol{\Psi}^{\top} \boldsymbol{\upsilon} + \boldsymbol{\varepsilon}$
- a quadratic penalty $\operatorname{pen}_G(\boldsymbol{\upsilon}) = \|G\boldsymbol{\upsilon}\|^2/2$.
- Penalized MLE: with $\mathbb{F}_G \stackrel{\text{def}}{=} \boldsymbol{\Psi}\boldsymbol{\Psi}^{\top} + G^2$

$$\widetilde{\boldsymbol{\upsilon}}_G = \mathbb{F}_G^{-1} \boldsymbol{\Psi} \boldsymbol{Y} ,$$

$$2L_G(\widetilde{\boldsymbol{\upsilon}}_G) - 2L_G(\boldsymbol{\upsilon}_G^*) = \|\mathbb{F}_G^{-1/2} \boldsymbol{\Psi} \boldsymbol{\varepsilon}\|^2 .$$

where $\boldsymbol{\upsilon}_G^* = \mathbb{F}_G^{-1} \boldsymbol{\Psi} \mathbb{E} \boldsymbol{Y}$.

Loss, bias-variance decomposition:

$$\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* = \mathbb{F}_G^{-1} \boldsymbol{\Psi} \boldsymbol{\varepsilon} + \mathbb{F}_G^{-1} G^2 \boldsymbol{\upsilon}^*,$$

$$\mathbb{E}\big\|Q(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*)\big\|^2 = \operatorname{tr}\{Q \mathbb{F}_G^{-1} \operatorname{Var}(\boldsymbol{\Psi}\boldsymbol{\varepsilon}) \mathbb{F}_G^{-1} Q^{\top}\} + \|Q \mathbb{F}_G^{-1} G^2 \boldsymbol{\upsilon}^*\|^2.$$

Stochastic component $\zeta(\boldsymbol{v}) \stackrel{\text{def}}{=} L(\boldsymbol{v}) - \mathbb{E}L(\boldsymbol{v})$ is linear in $\boldsymbol{v}$:

$$\nabla\zeta \stackrel{\text{def}}{=} \nabla\zeta(\boldsymbol{v});$$

The function $f(\boldsymbol{v}) = \mathbb{E}L(\boldsymbol{v})$ is smooth and concave in $\boldsymbol{v}$.

Consider

$$\widetilde{\boldsymbol{v}}_G = \operatorname*{argmax}_{\boldsymbol{v}} L_G(\boldsymbol{v}) = \operatorname*{argmax}_{\boldsymbol{v}} \Bigl\{ L(\boldsymbol{v}) - \frac{1}{2}\|G\boldsymbol{v}\|^2 \Bigr\};$$

$$\boldsymbol{v}_G^* = \operatorname*{argmax}_{\boldsymbol{v}} \mathbb{E}\, L_G(\boldsymbol{v}) = \operatorname*{argmax}_{\boldsymbol{v}} \Bigl\{ \mathbb{E}\, L(\boldsymbol{v}) - \frac{1}{2}\|G\boldsymbol{v}\|^2 \Bigr\};$$

$$\boldsymbol{v}^* = \operatorname*{argmax}_{\boldsymbol{v}} \mathbb{E}\, L(\boldsymbol{v});$$

$(\mathcal{C}_G)$ *The function* $\mathbb{E}L_G(\boldsymbol{\upsilon})$ *is concave on* Υ *which is open and convex set in* $\mathbb{R}^p$ *.*

(ζ) *The stochastic component* $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E}L(\boldsymbol{\upsilon})$ *is linear in* $\boldsymbol{\upsilon}$ *,* $\nabla\zeta \equiv \nabla\zeta(\boldsymbol{\upsilon}) \in \mathbb{R}^p$ *.*

$f(\boldsymbol{v}) = \mathbb{E} L_G(\boldsymbol{v})$ is smooth: for $k = 3$ (and may be $k = 4$)

$(\mathcal{T}_3^*)$ $f(\boldsymbol{v})$ *is strongly concave,* $\mathbb{D}^2 \leq \nabla^2 f(\boldsymbol{v})$ *, and*

$$\sup_{\boldsymbol{u} \colon \|\mathbb{D}\boldsymbol{u}\| \leq \mathtt{r}} \; \sup_{\boldsymbol{z} \in \mathbb{R}^p} \frac{\bigl| \langle \nabla^3 f(\boldsymbol{v} + \boldsymbol{u}), \boldsymbol{z}^{\otimes k} \rangle \bigr|}{\|\mathbb{D}\boldsymbol{z}\|^3} \leq \tau_3 \, .$$

Banach's characterization [Banach, 1938] yields for $k \geq 2$

$$\bigl| \langle \nabla^k f(\boldsymbol{v} + \boldsymbol{u}), \boldsymbol{z}_1 \otimes \ldots \otimes \boldsymbol{z}_k \rangle \bigr| \leq \tau_k \|\mathbb{D}\boldsymbol{z}_1\| \ldots \|\mathbb{D}\boldsymbol{z}_k\| \, .$$

If $f(\boldsymbol{v}) = \mathbb{E} L_G(\boldsymbol{v})$ scales with n, then the same holds for $\nabla^k f(\boldsymbol{v})$ and

$$\tau_3 \asymp n^{-1/2} , \qquad \tau_4 \asymp n^{-1} \, .$$

By $(\boldsymbol{\zeta})$, it holds for $\zeta(\boldsymbol{v}) = L(\boldsymbol{v}) - \mathbb{E}L(\boldsymbol{v})$

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

$(\boldsymbol{\nabla\zeta})$ *There exists* $V^2 \geq \operatorname{Var}(\nabla\zeta)$ *s.t.* $\boldsymbol{\xi} \stackrel{\text{def}}{=} V^{-1}\nabla\zeta$ *satisfies for any considered* $\mathtt{x} > 0$ *and* $B \in \mathfrak{M}_p$

$$\mathbb{P}\left(\|B^{1/2}\boldsymbol{\xi}\| \geq z(B,\mathtt{x})\right) \leq 3\mathrm{e}^{-\mathtt{x}},$$

$$z^2(B,\mathtt{x}) \stackrel{\text{def}}{=} \operatorname{tr} B + 2\sqrt{\mathtt{x}\ \operatorname{tr} B^2} + 2\mathtt{x}\|B\|\,.$$

Alternative formulation: on $\Omega(\mathtt{x})$ with $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - 3\mathrm{e}^{-\mathtt{x}}$

$$\|B^{1/2}\boldsymbol{\xi}\| \geq z(B,\mathtt{x}).$$

With the metric tensor D from $(\mathcal{T}_3^*)$, define

$$\mathtt{r}_D = z(B_D, \mathtt{x}), \quad B_D \stackrel{\text{def}}{=} \operatorname{Var}(D \mathbb{F}_G^{-1} \nabla \zeta), \quad \mathbb{F}_G = \mathbb{F}_G(\boldsymbol{\upsilon}_G^*).$$

Theorem (Fisher and Wilks expansions)

Assume $(\mathcal{C}_G)$, (ζ), $(\nabla\zeta)$, *and* $(\mathcal{T}_3^*)$ *with* D, $\mathtt{r}$, *and* τ_3 *s.t.*

$$D^2 \le \mathbb{F}_G, \quad \mathtt{r} \ge \frac{3}{2} \mathtt{r}_D, \quad \tau_3 \, \mathtt{r}_D < \frac{4}{9},$$

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

$$\bigl\| D^{-1} \mathbb{F}_G(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^* - \mathbb{F}_G^{-1} \nabla \zeta) \bigr\| \le \frac{3\tau_3}{4} \| D \mathbb{F}_G^{-1} \nabla \zeta \|^2,$$

$$\bigl| 2L_G(\widetilde{\boldsymbol{\upsilon}}_G) - 2L_G(\boldsymbol{\upsilon}_G^*) - \| \mathbb{F}_G^{-1/2} \nabla \zeta \|^2 \bigr| \le \tau_3 \| D \mathbb{F}_G^{-1} \nabla \zeta \|^3 .$$

Compare

$$\boldsymbol{\upsilon}_G^* = \operatorname{argmax}\Bigl\{ \mathbb{E}L(\boldsymbol{\upsilon}) - \frac{1}{2}\|G\boldsymbol{\upsilon}\|^2 \Bigr\}, \qquad \boldsymbol{\upsilon}^* = \operatorname{argmax} \mathbb{E}L(\boldsymbol{\upsilon}).$$

Proposition

Let

$$\mathsf{b}_G \stackrel{\text{def}}{=} \|D \mathbb{F}_G^{-1} G^2 \boldsymbol{\upsilon}^*\|.$$

Assume $(\boldsymbol{\mathcal{T}_3^*})$ *with* $\mathtt{r} = (3/2)\mathsf{b}_G$ *and let* $\tau_3\, \mathsf{b}_G \le 1/2$ *. Then*

$$\|D^{-1} \mathbb{F}_G (\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^* + \mathbb{F}_G^{-1} G^2 \boldsymbol{\upsilon}^*)\| \le \frac{3\tau_3}{4}\, \mathsf{b}_G^2 \,.$$

Theorem

For any linear Q

$$\|Q(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* - \mathbb{F}_G^{-1}\nabla\zeta + \mathbb{F}_G^{-1}G^2\boldsymbol{\upsilon}^*)\|$$

$$\leq \|Q\mathbb{F}_G^{-1}D\| \, \frac{3\tau_3}{4} \left(\|D\mathbb{F}_G^{-1}\nabla\zeta\|^2 + \mathsf{b}_D^2\right)$$

Fix $Q\colon \mathbb{R}^p \to \mathbb{R}^q$ and define

$$\mathrm{p}_D \stackrel{\text{def}}{=} \operatorname{tr}\operatorname{Var}(D \mathbb{F}_G^{-1} \nabla\zeta)\,, \qquad \mathsf{b}_D = \| D \mathbb{F}_G^{-1} G^2 \boldsymbol{\upsilon}^* \|\,,$$

$$\mathrm{p}_Q \stackrel{\text{def}}{=} \operatorname{tr}\operatorname{Var}(Q \mathbb{F}_G^{-1} \nabla\zeta)\,, \qquad \mathsf{b}_Q = \| Q \mathbb{F}_G^{-1} G^2 \boldsymbol{\upsilon}^* \|\,,$$

$$\mathscr{R}_Q \stackrel{\text{def}}{=} \mathbb{E}\bigl\{ \| Q \mathbb{F}_G^{-1} (\nabla\zeta - G^2 \boldsymbol{\upsilon}^*) \|^2 \, \mathbb{1}_{\Omega(\mathrm{x})} \bigr\} \le \mathrm{p}_Q + \mathsf{b}_Q^2\,.$$

Theorem

$$\mathbb{E}\bigl\{ \| Q(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*) \| \, \mathbb{1}_{\Omega(\mathrm{x})} \bigr\} \le \mathscr{R}_Q^{1/2} + \| Q \mathbb{F}_G^{-1} D \| \, \frac{3\tau_3}{4} (\mathrm{p}_D + \mathsf{b}_D^2)\,,$$

$$(1 - \alpha_Q)^2 \mathscr{R}_Q \le \mathbb{E}\bigl\{ \| Q\,(\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^*) \|^2 \, \mathbb{1}_{\Omega(\mathrm{x})} \bigr\} \le (1 + \alpha_Q)^2 \mathscr{R}_Q$$

provided $\quad \alpha_Q \stackrel{\text{def}}{=} \dfrac{\| Q \mathbb{F}_G^{-1} D \| \, (3/4) \tau_3 \, (\mathrm{p}_D + \mathsf{b}_D^2)}{\sqrt{\mathscr{R}_Q}} < 1\,.$

With $n = \lambda_{\min}(D^2)$, $Q = D = n^{1/2} I\!I_p$, and $\mathscr{R}_G = \mathrm{p}_G + \mathsf{b}_G^2$

$$\mathbb{E}\bigl\{ \| n^{1/2} \, (\widetilde{\boldsymbol{v}}_G - \boldsymbol{v}^*) \|^2 \, 1\!\!I_{\Omega(\mathtt{x})} \bigr\} = \mathscr{R}_G \bigl(1 \pm \tau_3 \sqrt{\mathscr{R}_G} \bigr).$$

A sharp bound under $\tau_3 \sqrt{\mathrm{p}_G} \ll 1$ and $\tau_3 \, \mathsf{b}_G \ll 1$.

Critical dimension: with $\tau_3 \asymp n^{-1/2}$

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

Let observations $Y_1, \ldots, Y_n$ follow the nonlinear regression model

$$Y_i = m(\boldsymbol{X}_i, \boldsymbol{\theta}) + \varepsilon_i$$

with independent zero mean errors ε_i .

Target parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ for p large/infinite.

Example in mind: $\boldsymbol{\theta}$ codes the architecture of a DNN.

Aim: estimation/inference on $\boldsymbol{\theta}$.

Least squares estimation (Gauss, Legendre):

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{Y} - m(\boldsymbol{X}_i, \boldsymbol{\theta})\|^2.$$

Problems: $L(\boldsymbol{\theta})$ is not concave, the gradient $\nabla\zeta(\boldsymbol{\theta}) = \nabla m(\boldsymbol{\theta})\boldsymbol{\varepsilon}$ of the stochastic component depends on $\boldsymbol{\theta}$, both SLS assumptions fail.

Calming = (pre)smoothing + relaxation + regularization.

(Pre)smoothing (or dual representation/kernelization/observables):

$$\boldsymbol{Z} = \boldsymbol{\Phi}\,\boldsymbol{Y}, \quad \boldsymbol{\Phi} \colon \mathbb{R}^n \to \mathbb{R}^q .$$

Further, define $\boldsymbol{M}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \boldsymbol{\Phi}\,\boldsymbol{m}(\boldsymbol{\theta})$ and represent

$$\boldsymbol{Y} = \boldsymbol{m}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \quad \to \quad \boldsymbol{\Phi}\,\boldsymbol{Y} \approx \boldsymbol{\eta} + \boldsymbol{\Phi}\boldsymbol{\varepsilon} \quad \text{and} \quad \boldsymbol{\eta} \approx \boldsymbol{M}(\boldsymbol{\theta}).$$

Then $\|\boldsymbol{Y} - \boldsymbol{m}(\boldsymbol{\theta})\|^2$ transforms to

$$\|\boldsymbol{\Phi}\boldsymbol{Y} - \boldsymbol{\eta}\|^2 + \lambda\|\boldsymbol{\Phi}\,\boldsymbol{m}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 = \|\boldsymbol{Z} - \boldsymbol{\eta}\|^2 + \lambda\|\boldsymbol{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2$$

with a Lagrange multiplier λ. Leads to

$$2\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\|\boldsymbol{Z} - \boldsymbol{\eta}\|^2 - \lambda\|\boldsymbol{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 ,$$

$$2\mathscr{L}_{\mathcal{G}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\|\boldsymbol{Z} - \boldsymbol{\eta}\|^2 - \lambda\|\boldsymbol{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 - \|G\boldsymbol{\theta}\|^2 - \|\varGamma\boldsymbol{\eta}\|^2$$

Consider (with $\lambda = 1$)

$$\begin{aligned} \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) &= -\frac{1}{2}\|\mathcal{S}\boldsymbol{Y} - \boldsymbol{\eta}\|^2 - \frac{1}{2}\|\mathcal{S}\,\boldsymbol{m}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 \\ &= -\frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{\eta}\|^2 - \frac{1}{2}\|\boldsymbol{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 \end{aligned}$$

$\widetilde{\boldsymbol{\upsilon}}_G$ is given by

$$\begin{aligned} \mathscr{L}_G(\boldsymbol{\upsilon}) &= \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{\eta}\|^2 - \frac{1}{2}\|\boldsymbol{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 - \frac{1}{2}\|G\boldsymbol{\theta}\|^2, \\ \widetilde{\boldsymbol{\upsilon}}_G &= \operatorname*{argmax}_{\boldsymbol{\upsilon}\in\Upsilon} \mathscr{L}_G(\boldsymbol{\upsilon}). \end{aligned}$$

Profile MLE: $\quad \widetilde{\boldsymbol{\theta}}_G = \operatorname*{argmax}_{\boldsymbol{\theta}} \max_{\boldsymbol{\eta}} \mathscr{L}_G(\boldsymbol{\upsilon}).$

With $\boldsymbol{m}^* = \mathbb{E}\boldsymbol{Y}$ and $\boldsymbol{M}^* = \mathcal{S}\boldsymbol{m}^*$

$$\boldsymbol{\upsilon}^* = \operatorname*{argmin}_{\boldsymbol{\upsilon}=(\boldsymbol{\theta},\boldsymbol{\eta})\in\Upsilon} \left\{ \|\boldsymbol{M}^* - \boldsymbol{\eta}\|^2 + \|\boldsymbol{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 \right\},$$

$$\boldsymbol{\upsilon}^*_G = \operatorname*{argmin}_{\boldsymbol{\upsilon}=(\boldsymbol{\theta},\boldsymbol{\eta})\in\Upsilon} \left\{ \|\boldsymbol{M}^* - \boldsymbol{\eta}\|^2 + \|\boldsymbol{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 + \|G\boldsymbol{\theta}\|^2 \right\}.$$

The $\boldsymbol{\theta}$-component $\boldsymbol{\theta}^*$ of $\boldsymbol{\upsilon}^*$ (resp. $\boldsymbol{\theta}^*_G$ of $\boldsymbol{\upsilon}^*_G$) solves the original problem in which the smoothed response $\boldsymbol{Z} = \mathcal{S}\,\boldsymbol{Y}$ is replaced by the auxiliary parameter $\boldsymbol{\eta}^*$ (resp. $\boldsymbol{\eta}^*_G$):

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}^*_G\|^2,$$

$$\boldsymbol{\theta}^*_G = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \left\{ \|\boldsymbol{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}^*_G\|^2 + \|G\boldsymbol{\theta}\|^2 \right\}.$$

Let

$$D^2(\boldsymbol{\theta}) = \frac{1}{2} \nabla \boldsymbol{M}(\boldsymbol{\theta})\ \nabla \boldsymbol{M}(\boldsymbol{\theta})^\top = \frac{1}{2} \sum_{j=1}^{q} \nabla M_j(\boldsymbol{\theta})\ \nabla M_j(\boldsymbol{\theta})^\top \in \mathfrak{M}_p\,.$$

For an initial guess $\boldsymbol{\theta}_0$, define $D_0 = D(\boldsymbol{\theta}_0)$ and

$$\varTheta^\circ = \bigl\{\boldsymbol{\theta} \colon \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\| \le \mathtt{r}_0\bigr\}$$

$(\boldsymbol{\theta}^*)$ *It holds* $\boldsymbol{\theta}^* \in \varTheta^\circ$ *and* $\boldsymbol{\theta}^*_G \in \varTheta^\circ$ *.*

Conditions of this kind are often applied in nonlinear optimization for studying, e.g. Gauss-Newton iterations; see e.g. [Gratton et al., 2007].

With

$$D^2(\boldsymbol{\theta}) = \frac{1}{2} \nabla \boldsymbol{M}(\boldsymbol{\theta}) \, \nabla \boldsymbol{M}(\boldsymbol{\theta})^\top , \quad D_0 = D(\boldsymbol{\theta}_0) \, ,$$

assume

(∇M) *For some* $\omega^+ \le 1/3$ *and any* $\boldsymbol{\theta} \in \varTheta^\circ$ *, it holds*

$$(1 - \omega^+) \, D_0^2 \le D^2(\boldsymbol{\theta}) \le (1 + \omega^+) \, D_0^2.$$

$(\nabla^k M)$ *For* $k \in \{2, 3, 4\}$ *and small* $\varkappa \ge 0$ *, uniformly over* $\boldsymbol{\theta} \in \varTheta^\circ$ *and* $\boldsymbol{u} \in \mathbb{R}^p$

$$\sum_{j=1}^{q} \langle \nabla^k M_j(\boldsymbol{\theta}), \boldsymbol{u}^{\otimes k} \rangle^2 \le \varkappa^{2k-2} \, \| D_0 \boldsymbol{u} \|^{2k} \, .$$

For $\zeta(\boldsymbol{\upsilon}^*) = \mathscr{L}(\boldsymbol{\upsilon}) - \mathbb{E}\mathscr{L}(\boldsymbol{\upsilon})$, it holds

$$\nabla\zeta = \begin{pmatrix} 0 \\ \nabla_{\boldsymbol{\eta}}\zeta \end{pmatrix} = \begin{pmatrix} 0 \\ \mathcal{S}\boldsymbol{\varepsilon} \end{pmatrix},$$

Bounding $\nabla\zeta$ can be easily reduced to a similar question for $\mathcal{S}\boldsymbol{\varepsilon}$.

$(\mathcal{S}\boldsymbol{\varepsilon})$ *The vector $\mathcal{S}\boldsymbol{\varepsilon}$ satisfies for all considered $\mathtt{x} > 0$*

$$\mathbb{P}\bigl(\|\mathcal{S}\boldsymbol{\varepsilon}\| > z(\mathbb{V}^2, \mathtt{x})\bigr) \le 3\mathrm{e}^{-\mathtt{x}},$$

where

$$\mathbb{V}^2 \stackrel{\text{def}}{=} \operatorname{Var}(\mathcal{S}\boldsymbol{\varepsilon}) = \mathcal{S}\operatorname{Var}(\boldsymbol{\varepsilon})\,\mathcal{S}^\top,$$

$$z(\mathbb{V}^2, \mathtt{x}) \stackrel{\text{def}}{=} \sqrt{\operatorname{tr}\mathbb{V}^2} + \sqrt{2\mathtt{x}\,\|\mathbb{V}^2\|}\,.$$

[Spokoiny, 2024b], [Spokoiny, 2024a].

With

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{\eta}\|^2 - \frac{1}{2}\|\boldsymbol{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2$$

it holds

$$\mathscr{F}_G(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} -\nabla^2 \mathscr{L}_G(\boldsymbol{\upsilon}) = \begin{pmatrix} \mathbb{F}_G(\boldsymbol{\upsilon}) & -\nabla \boldsymbol{M}(\boldsymbol{\theta}) \\ -\nabla \boldsymbol{M}(\boldsymbol{\theta})^\top & 2\, \mathbb{I}_q \end{pmatrix}$$

with the upper left diagonal block

$$\mathbb{F}_G(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} \nabla \boldsymbol{M}(\boldsymbol{\theta})\, \nabla \boldsymbol{M}(\boldsymbol{\theta})^\top + \sum_{j=1}^{q} \{M_j(\boldsymbol{\theta}) - \eta_j\}\, \nabla^2 M_j(\boldsymbol{\theta}) + G^2 \,.$$

Define $\mathscr{F} \stackrel{\text{def}}{=} \mathscr{F}(\boldsymbol{\upsilon}_G^*)$.

With

$$D^2(\boldsymbol{\theta}) = \frac{1}{2}\nabla \boldsymbol{M}(\boldsymbol{\theta})\ \nabla \boldsymbol{M}(\boldsymbol{\theta})^\top$$

and $D^2 = D^2(\boldsymbol{\theta}_G^*)$, define

$$\mathcal{D}^2 = \operatorname{block}\{D^2, I\!I_q\}\,.$$

Lemma

It holds

$$\mathscr{F}_G^{-1} \le 2 \begin{pmatrix} (D^2 + 2G^2)^{-1} & 0 \\ 0 & I\!I_q \end{pmatrix}.$$

With $\boldsymbol{M}_G = (G^2 \boldsymbol{\theta}^*, 0)$, define

$$\mathsf{b}_{\mathcal{D}} \stackrel{\text{def}}{=} \|\mathcal{D}\, \mathscr{F}_G^{-1} \boldsymbol{M}_G\| \leq 2 \, \|D\, \mathbb{F}_G^{-1} G^2 \boldsymbol{\theta}^*\| \, .$$

Theorem

Let $\mathtt{r}_{\mathcal{D}} \stackrel{\text{def}}{=} 2 z(\mathbb{V}^2, \mathtt{x})$ *and*

$$\mathtt{r}_0 \, \varkappa < \frac{1}{4} , \quad \mathtt{r}_0 \geq \frac{3}{2} (\mathtt{r}_{\mathcal{D}} \vee \mathsf{b}_{\mathcal{D}}) , \quad \tau_3 (\mathtt{r}_{\mathcal{D}} \vee \mathsf{b}_{\mathcal{D}}) < \frac{2}{9} \, .$$

It holds on $\Omega(\mathtt{x})$ *with for any linear mapping* Q *on* $\boldsymbol{\theta}$

$$\left\| Q \{ \widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^* - (\mathscr{F}_G^{-1} \nabla \zeta)_{\boldsymbol{\theta}} + (\mathscr{F}_G^{-1} \boldsymbol{M}_G)_{\boldsymbol{\theta}} \} \right\|$$

$$\leq \| Q \, D^{-1} \| \, \frac{3 \tau_3}{2} \left(\| 2 \mathcal{S} \boldsymbol{\varepsilon} \|^2 + \mathsf{b}_{\mathcal{D}}^2 \right) .$$

Also, define

$$\mathrm{p}_Q \stackrel{\text{def}}{=} \operatorname{tr} \operatorname{Var}\{Q(\mathscr{F}_{\mathcal{G}}^{-1} \nabla \zeta)_{\boldsymbol{\theta}}\},$$

$$\mathscr{R}_Q \stackrel{\text{def}}{=} \mathbb{E}\{\|Q(\mathscr{F}_{\mathcal{G}}^{-1} \nabla \zeta)_{\boldsymbol{\theta}} - Q(\mathscr{F}_{\mathcal{G}}^{-1} \boldsymbol{M}_{\mathcal{G}})_{\boldsymbol{\theta}}\|^2 \, \mathbb{1}_{\Omega(\mathtt{x})}\}$$

$$\leq \mathrm{p}_Q + \|Q\,(\mathscr{F}_{\mathcal{G}}^{-1} \, \boldsymbol{M}_G)_{\boldsymbol{\theta}}\|^2 \, .$$

Theorem

With $\bar{\mathrm{p}}_{\mathcal{D}} = \mathbb{E}\|\mathcal{D}\, \mathscr{F}_{\mathcal{G}}^{-1} \nabla \zeta\|^2 \leq \mathbb{E}\|2\mathcal{S}\boldsymbol{\varepsilon}\|^2$ *, it holds*

$$\mathbb{E}\left\{\|Q(\widetilde{\boldsymbol{\theta}}_{\mathcal{G}} - \boldsymbol{\theta}^*)\| \, \mathbb{1}_{\Omega(\mathtt{x})}\right\} \leq \sqrt{\mathscr{R}_Q} + \|Q\, D^{-1}\| \, \frac{3\tau_3}{2} \left(\bar{\mathrm{p}}_{\mathcal{D}} + \mathsf{b}_{\mathcal{D}}^2\right) .$$

Define the full effective dimension

$$\bar{\mathrm{p}}_{\mathcal{D}} = \mathbb{E}\|2\mathcal{S}\boldsymbol{\varepsilon}\|^2 \leq 4\sigma^2 q$$

The *effective sample size* n is defined via the constant $\varkappa$ from $(\boldsymbol{\nabla}^{\boldsymbol{k}}\boldsymbol{M})$. We use

$$\tau_3 \asymp \varkappa \asymp n^{-1/2}.$$

The results require

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

Suppose that a matrix $\boldsymbol{Y} = (Y_{ij}) \in \mathbb{R}^{p \times q}$ is partly observed with noise:

$$Y_{ij} = X_{ij} + \varepsilon_{ij}\,, \quad (i,j) \in \mathcal{G},$$

where $\mathcal{G}$ describes the “design”. The goal is to recover the matrix $\boldsymbol{X} = (X_{ij})$ under a “low-rank” condition. The latter yields the representation

$$\boldsymbol{X} = \boldsymbol{U}\,\boldsymbol{\Lambda}\,\boldsymbol{V}^{\top} = \sum_{m} \lambda_m\,\boldsymbol{u}_m\,\boldsymbol{v}_m^{\top}\,, \tag{1}$$

where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$, $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r) \in \mathbb{R}^{p \times r}$, $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r) \in \mathbb{R}^{p \times r}$, and the vectors $\boldsymbol{u}_m$ are orthonormal in $\mathbb{R}^p$ while $\boldsymbol{v}_m$ are orthonormal in $\mathbb{R}^q$. In the case when all the eigenvalues λ_j are different and ordered by absolute values $|\lambda_1| > \ldots > |\lambda_r|$, representation (1) is unique.

Let now $(\mathcal{T}_k)$ be a collection of “templates” in $\mathbb{R}^{p\times q}$, $k = 1,\ldots,K$. A typical example is of the form

$$\mathcal{T} = \operatorname{diag}(\delta_i)\,\mathbf{1}_{p\times q}\,\operatorname{diag}(\delta'_j),$$

where $\mathbf{1}_{p\times q}$ is the matrix of ones in $\mathbb{R}^{p\times q}$ and $(\delta_1,\ldots,\delta_p)$, $(\delta'_1,\ldots,\delta'_q)$ are obtained as independent Bernoulli r.v.'s. Informally, we include in the template $\mathcal{T}$ each row i with probability $\alpha_{1,i}$ and each column j with probability $\alpha_{2,j}$. Define

$$z(\mathcal{T}) \stackrel{\text{def}}{=} \langle \boldsymbol{X}, \mathcal{T} \rangle = \sum_{(i,j)\in\mathcal{G}} \mathcal{T}_{ij}\,X_{ij}$$

$$Z(\mathcal{T}) \stackrel{\text{def}}{=} \langle \boldsymbol{Y}, \mathcal{T} \rangle = \sum_{(i,j)\in\mathcal{G}} \mathcal{T}_{ij}\,Y_{ij}\,.$$

Also introduces “observables” Z_k and the image parameters z_k

$$Z_k = \langle \boldsymbol{Y}, \mathcal{T}_k \rangle, \qquad z_k = \langle \boldsymbol{X}, \mathcal{T}_k \rangle.$$

The whole set of parameters include orthonormal vectors $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$ in $\mathbb{R}^p$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r)$ in $\mathbb{R}^q$, the vector of eigenvalues $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^\top$, and the image vector $\boldsymbol{z} = (z_k)$ leading to the log-likelihood

$$\begin{aligned}\mathscr{L}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\lambda}, \boldsymbol{z}) &= -\frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{z}\|^2 - \frac{1}{2}\sum_{k=1}^{K}\bigl|z_k - \langle \mathcal{T}_k, \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top\rangle\bigr|^2 \\ &= -\frac{1}{2}\sum_{k=1}^{K}|\langle \boldsymbol{Y}, \mathcal{T}_k\rangle - z_k|^2 - \frac{1}{2}\sum_{k=1}^{K}\Bigl|z_k - \sum_{m=1}^{r}\lambda_m\,\boldsymbol{u}_m^\top \mathcal{T}_k\,\boldsymbol{v}_m\Bigr|^2 .\end{aligned}$$

(Near) orthonormality of the $\boldsymbol{u}_m$'s and $\boldsymbol{v}_m$'s can be enforced by the penalty

$$\mu\bigl(\|\boldsymbol{U}^\top\boldsymbol{U} - I\!I_r\|_{\mathrm{Fr}}^2 + \|\boldsymbol{V}^\top\boldsymbol{V} - I\!I_r\|_{\mathrm{Fr}}^2\bigr) .$$

Identifiability of the model is supported by a penalty $\sum_m g_m^2 \lambda_m^2$ on the eigenvalues $\lambda_1, \ldots, \lambda_r$ with $g_1^2 < \ldots < g_r^2$. Finally, to distinguish between $\boldsymbol{u}_m$ and $-\boldsymbol{u}_m$, add the penalty $\|\boldsymbol{U} - \boldsymbol{E}\|_{\mathrm{Fr}}^2 = \sum_m \|\boldsymbol{u}_m - \boldsymbol{e}_m\|^2$ for given orthonormal vectors $\boldsymbol{e}_m$ and similarly for $\boldsymbol{v}_m$.

In total

$$\mathscr{L}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\lambda}, \boldsymbol{z}) = -\frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{z}\|^2 - \frac{1}{2}\sum_{k=1}^{K} \bigl| \boldsymbol{z}_k - \langle \mathcal{T}_k, \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top \rangle \bigr|^2$$

$$- \frac{\mu_g}{2} \sum_{m=1}^{r} g_m^2 \lambda_m^2$$

$$- \frac{\mu_o}{2} \bigl(\|\boldsymbol{U}^\top \boldsymbol{U} - I\!I_r\|_{\mathrm{Fr}}^2 + \|\boldsymbol{V}^\top \boldsymbol{V} - I\!I_r\|_{\mathrm{Fr}}^2 \bigr)$$

$$- \frac{\mu_e}{2} \bigl(\|\boldsymbol{U} - \boldsymbol{E}\|_{\mathrm{Fr}}^2 + \|\boldsymbol{V} - \boldsymbol{E}'\|_{\mathrm{Fr}}^2 \bigr) . \tag{2}$$

The whole procedure includes the following steps:

- Fix a collection of templates $\mathcal{T}_k$ and compute $Z_k = \langle \mathcal{T}_k, \boldsymbol{Y} \rangle$;
- Fix the matrices $\boldsymbol{E} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_r) \in \mathbb{R}^{p\times r}$ with $\boldsymbol{E}^\top \boldsymbol{E} = \mathbb{I}_r$ and similarly $\boldsymbol{E}' = (\boldsymbol{e}_1', \ldots, \boldsymbol{e}_r') \in \mathbb{R}^{q\times r}$ with ${\boldsymbol{E}'}^\top \boldsymbol{E}' = \mathbb{I}_r$
- Solve the maximization problem $\widetilde{\boldsymbol{\upsilon}} = \operatorname{argmax}_{\boldsymbol{\upsilon}} \mathscr{L}(\boldsymbol{\upsilon})$ for $\mathscr{L}(\boldsymbol{\upsilon})$ from (2) by alternating optimization.
- Build $\widetilde{\boldsymbol{X}}$ using the solution $\widetilde{\boldsymbol{\upsilon}}$ e.g.

$$\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{U}} \operatorname{diag}(\widetilde{\boldsymbol{\lambda}}) \, \widetilde{\boldsymbol{V}}^\top = \sum_{m=1}^{r} \widetilde{\lambda}_m \, \widetilde{\boldsymbol{u}}_m \, \widetilde{\boldsymbol{v}}_m^\top .$$

- If necessary, redesign $\boldsymbol{E}$ and $\boldsymbol{E}'$ and repeat.

With $\phi(\boldsymbol{x}) = \mathtt{C} \exp(-\|\boldsymbol{x}\|^2/2)$, consider

$$f(\boldsymbol{x}) = \int \phi\Bigl(\frac{\boldsymbol{x}-\boldsymbol{m}}{\sigma}\Bigr)\, d\mu(\boldsymbol{m},\sigma)\,. \tag{3}$$

Usually, the mixing measure μ is discrete with well separated atoms $\{(\boldsymbol{m}_k,\sigma_k), k \in \mathscr{K}\}$:

$$\mu = \sum_{k\in\mathscr{K}} \mu_k \, \mathbb{I}_{\boldsymbol{m}_k,\sigma_k} \ .$$

Then $\mu = \mu_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} = \bigl\{(\mu_k,\boldsymbol{m}_k,\sigma_k), k \in \mathscr{K}\bigr\}$, where $\sum_k \mu_k = 1$.

$$f(\boldsymbol{x}) = f(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{k\in\mathscr{K}} \mu_k \, \phi\Bigl(\frac{\boldsymbol{x}-\boldsymbol{m}_k}{\sigma_k}\Bigr) .$$

Later we consider $\mu = \mu_{\boldsymbol{\theta}} = \sum_j \mu_k \, \delta_{\boldsymbol{m}_k,\sigma_k}$.

Let X_i be i.i.d. from f. Consider the problem of recovering the mixing measure μ from the data. Given a family of test functions $(\psi_j(\boldsymbol{x}))$, define the observables

$$Z_j \stackrel{\mathrm{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \psi_j(\boldsymbol{X}_i) \, .$$

One example is given by a collection $\psi_j(\boldsymbol{x}) = \psi(\|\boldsymbol{x} - \boldsymbol{x}_j\|^2 / s_j^2)$ for a kernel ψ, a fixed set of points $\boldsymbol{x}_j$ and scalings s_j, $j \leq q$. Denote

$$\psi_j(\boldsymbol{m}, \sigma) \stackrel{\mathrm{def}}{=} \int \psi_j(\boldsymbol{x}) \, \phi\Bigl(\frac{\boldsymbol{x} - \boldsymbol{m}}{\sigma}\Bigr) \, d\boldsymbol{x}$$

Under (3), it holds

$$\begin{aligned} \mathbb{E} Z_j &= \int \psi_j(\boldsymbol{x}) \, f(\boldsymbol{x}) d\boldsymbol{x} = \iint \psi_j(\boldsymbol{x}) \, \phi\Bigl(\frac{\boldsymbol{x} - \boldsymbol{m}}{\sigma}\Bigr) \, d\mu(\boldsymbol{m}, \sigma) \, d\boldsymbol{x} \\ &= \int \psi_j(\boldsymbol{m}, \sigma) \, d\mu(\boldsymbol{m}, \sigma) = \mu(\psi_j). \end{aligned}$$

The calming device suggests to introduce the image parameter $\boldsymbol{z}$ and the extended log-likelihood $\mathscr{L}(\boldsymbol{\upsilon}) = \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{z})$ with

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{z}) \stackrel{\text{def}}{=} -\frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{z}\|^2 - \frac{1}{2}\|\boldsymbol{z} - \mu_{\boldsymbol{\theta}}(\boldsymbol{\Psi})\|^2 ,$$

where $\mu_{\boldsymbol{\theta}}(\boldsymbol{\Psi})$ is a vector in $\mathbb{R}^q$ with the entries

$$\mu_{\boldsymbol{\theta}}(\psi_j) = \sum_{k \in \mathscr{K}} \mu_k \, \psi_j(\boldsymbol{m}_k, \sigma_k) \, .$$

To overcome the issue of identifiability problem, introduce a penalty

$$\mathrm{pen}_{\varkappa}(\boldsymbol{\mu}) \stackrel{\mathrm{def}}{=} \sum_{k \in \mathscr{K}} \varkappa_k^2 \, \mu_k^2 \, ,$$

where $\varkappa_k^2$ strictly increase with k. Such a penalty ensures identifiability if the true weights μ_k^* of each component $\phi_{\boldsymbol{m}_k, \sigma_k}$ are significantly different. Unfortunately, the problem is not completely resolved if there are different components with nearly the same weights μ_k^*. One possibility to make it fixed is by using an additional penalty $\mathrm{pen}_G(\boldsymbol{\theta})$ based on the distance of each mean $\boldsymbol{m}_k$ from the origin (or any other fixed point $\boldsymbol{m}_0$):

$$\mathrm{pen}_G(\boldsymbol{m}) \stackrel{\mathrm{def}}{=} \sum_{k \in \mathscr{K}} \| G \boldsymbol{m}_k \|^2 \, ,$$

where the matrix G identifies the distance from each mean $\boldsymbol{m}_k$ to the origin. In particular, one can use $G^2 = \mathrm{diag}(g_j^2)$ with g_j^2 strictly increasing.

Altogether leads to the following approach: for $\boldsymbol{\upsilon} = (\boldsymbol{\theta}, \boldsymbol{z}) = \{(\mu_k, \boldsymbol{m}_k, \sigma_k), (z_j)\}$ and $\operatorname{pen}(\boldsymbol{\upsilon}) = \operatorname{pen}_{\varkappa}(\boldsymbol{\theta}) + \operatorname{pen}_G(\boldsymbol{m})$

$$\begin{aligned}\widetilde{\boldsymbol{\upsilon}}_{\mathcal{G}} &= \operatorname*{argmax}_{\boldsymbol{\upsilon}} \Bigl\{ \mathscr{L}(\boldsymbol{\upsilon}) - \frac{1}{2} \operatorname{pen}(\boldsymbol{\upsilon}) \Bigr\} \\ &= \operatorname*{argmax}_{\boldsymbol{\upsilon}} \biggl\{ -\frac{1}{2} \| \boldsymbol{Z} - \boldsymbol{z} \|^2 - \frac{1}{2} \| \boldsymbol{z} - \mu_{\boldsymbol{\theta}}(\boldsymbol{\Psi}) \|^2 \\ &\qquad - \frac{1}{2} \sum_{k \in \mathscr{K}} \| G \boldsymbol{m}_k \|^2 - \frac{1}{2} \sum_{k \in \mathscr{K}} \varkappa_k^2 \, \mu_k^2 \biggr\}.\end{aligned}$$

The structural penalty $\| \boldsymbol{z} - \mu_{\boldsymbol{\theta}}(\boldsymbol{\Psi}) \|^2$ creates some difficulties for the analysis, however, it is deterministic and smooth in the scope of arguments.

Let, for an input vector $\boldsymbol{x} = (x_m) \in \mathbb{R}^d$, the hidden layer transformation is given by

$$\boldsymbol{x}^{(1)} = \sigma(\boldsymbol{a} + W\boldsymbol{x}),$$

where $\boldsymbol{a} \in \mathbb{R}^p$, $W : \mathbb{R}^d \to \mathbb{R}^p$, and σ is a coordinate-wise activating function, e.g.

$$\sigma(t) = \lambda^{-1} \log(1 + \mathrm{e}^{\lambda t}).$$

The transformed vectors $\boldsymbol{x}^{(1)}$ enter in the logistic regression model for binary labels Y_i

$$\mathbb{P}(Y = 1 \,|\, \boldsymbol{x}^{(1)}) = \mathrm{softmax}(\boldsymbol{x}^{(1)}), \qquad \mathbb{P}(Y = 0 \,|\, \boldsymbol{x}^{(1)}) = 1 - \mathrm{softmax}(\boldsymbol{x}^{(1)}) \,.$$

The structure of this neuronal network is described by the structural parameter $\boldsymbol{\upsilon} = (\boldsymbol{a}, W)$.

Now consider the statistical problem of inference about this parameter given independent data $(\boldsymbol{X}_i, Y_i)$. The corresponding log-likelihood involves the fidelity term $L(\boldsymbol{Y}, \boldsymbol{x}^{(1)}) = \sum_i \ell(Y_i, \eta_i)$ with $\ell(y, \eta) = y\eta - \log(1 + \mathrm{e}^{\eta})$, $\eta_i = \operatorname{softmax}(\boldsymbol{x}_i^{(1)})$ and the structural terms $\|\boldsymbol{X}^{(1)} - \sigma(\boldsymbol{a} + W\boldsymbol{X})\|^2$. We also add some penalty on $\boldsymbol{a} = (a_j)$ and $W = (w_{mj})$:

$$\operatorname{pen}(\boldsymbol{a}) = \frac{1}{2}\|\mathcal{T}\boldsymbol{a}\|^2 = \frac{1}{2}\sum_j a_j^2\,\mathcal{T}_j^2\,,$$

$$\operatorname{pen}(W) = \frac{1}{2}\langle \mathcal{G}, W\rangle^2 = \frac{1}{2}\sum_{m,j} w_{mj}^2 \mathcal{G}_{mj}^2\,,$$

with $\mathcal{T}_j$ and $\mathcal{G}_{mj}$ polynomially growing in j. This enables us to identify the most informative nodes in the hidden layer and control the overall complexity of the network.

This results in maximization of the penalized log-likelihood

$$\mathscr{L}(\boldsymbol{a}, W, \boldsymbol{X}^{(1)}) = L(\boldsymbol{Y}, \mathrm{softmax}(\boldsymbol{X}^{(1)})) - \frac{\mu}{2}\|\boldsymbol{X}^{(1)} - \sigma(\boldsymbol{a} + W\boldsymbol{X})\|^2$$

$$-\frac{1}{2}\|\mathcal{T}\boldsymbol{a}\|^2 - \frac{1}{2}\langle\mathcal{G}, W\rangle^2$$

with a Lagrange multiplyer μ. The structural relation $\boldsymbol{X}^{(1)} \equiv \sigma(\boldsymbol{a} + W\boldsymbol{X})$ is relaxed and replaced by the structural penalty $\frac{\mu}{2}\|\boldsymbol{X}^{(1)} - \sigma(\boldsymbol{a} + W\boldsymbol{X})\|^2$. Introducing the auxillary variable $\boldsymbol{X}^{(1)}$ is not mandatory, one can use $\boldsymbol{X}^{(1)} \equiv \sigma(\boldsymbol{a} + W\boldsymbol{X})$. However, it can be useful, e.g. for an additional penalization.

One example of choosing the penalty on $\boldsymbol{a}$ and W is given by $\mathcal{T}_j^2 = \mathsf{c}_a j^{2\beta}$, and $\mathcal{G}_{mj}^2 = \mathcal{G}_j^2 = \mathsf{c}_w j^{2\beta}$ for e.g. $\beta = 2$ and some constants $\mathsf{c}_a, \mathsf{c}_w$. Any prior information about the input features $\boldsymbol{X}$ can be incorporated in the penalty coefficients $\mathcal{A}_m$ leading to a structure $\mathcal{G}_{mj}^2 = G_m^2 + \mathcal{G}_j^2$, e.g. $\mathcal{G}_{mj}^2 = \mathsf{c}_x m^{2\beta} + \mathsf{c}_w j^{2\beta}$.

This construction extends to a K -layer network using recurrence

$$\boldsymbol{X}^{(k)} = \sigma(\boldsymbol{a}^{(k)} + W^{(k)}\boldsymbol{X}^{(k-1)})$$

for $k = 1, \ldots, K$ and $\boldsymbol{X}^{(0)} = \boldsymbol{X}$. This leads to the log-likelihood

$$\mathscr{L}_{\mathcal{G}}(\boldsymbol{X}^{(1)}, \boldsymbol{a}^{(1)}, W^{(1)}, \ldots, \boldsymbol{X}^{(K)}, \boldsymbol{a}^{(K)}, W^{(K)}) = L(\boldsymbol{Y}, \boldsymbol{X}^{(K)})$$

$$- \frac{1}{2} \sum_{k=1}^{K} \Big(\|\boldsymbol{X}^{(k)} - \sigma(\boldsymbol{a}^{(k)} + W^{(k)}\boldsymbol{X}^{(k-1)})\|^2 + \|\mathcal{T}^{(k)}\boldsymbol{a}^{(k)}\|^2 + \langle W^{(k)}, \mathcal{G}^{(k)} \rangle^2 \Big)$$

Let $\mathbb{P}_f \sim \exp f(\boldsymbol{x})$. Denote by $\mathbb{N}_{\boldsymbol{x},\mathbb{Z}}$ the Gaussian measure with the mean $\boldsymbol{x}$ and covariance $\mathbb{Z}^{-1}$, i.e. $\mathbb{N}_{\boldsymbol{x},\mathbb{Z}} \stackrel{\text{def}}{=} \mathcal{N}(\boldsymbol{x}, \mathbb{Z}^{-1})$.

$$\text{Gauss VI:} \quad (\boldsymbol{x}_{\text{VI}}, \mathbb{Z}_{\text{VI}}) = \operatorname*{arginf}_{\boldsymbol{x},\mathbb{Z}} \mathscr{K}(\mathbb{N}_{\boldsymbol{x},\mathbb{Z}} \,\|\, \mathbb{P}_f).$$

Natural candidates:

1. Laplace: $\boldsymbol{x}_{\text{VI}} \approx \operatorname{argmax} f(\boldsymbol{x})$, $\mathbb{Z}_{\text{VI}} \approx -\nabla^2 f(\boldsymbol{x}^*)$;
2. Moments: $\boldsymbol{x}_{\text{VI}} \approx \mathbb{E}_f \boldsymbol{X}$, $\mathbb{Z}_{\text{VI}}^{-1} \approx \operatorname{Var}_f(\boldsymbol{X})$.

[Katsevich and Rigollet, 2023] argued for (2).

- [David M. Blei and McAuliffe, 2017] Variational Inference: A review for statisticians
- [Zhang and Gao, 2020] Convergence rates of variational posterior distributions
- [Wang and Blei, 2019] Frequentist consistency of variational Bayes
- [Han and Yang, 2019] Statistical inference in mean-field variational Bayes
- [Challis and Barber, 2013] Gaussian Kullback-Leibler approximate inference
- [Alquier and Ridgway, 2020] Concentration of tempered posteriors and of their variational approximations
- [Lambert et al., 2023] Variational inference via Wasserstein gradient flows

The VI approach assumes minimizing of the KL-divergence $\mathscr{K}(\mathbb{N}_{\boldsymbol{x},\mathbb{Z}} \,\|\, \mathbb{P}_f)$ over all feasible $\boldsymbol{x}, \mathbb{Z}$. Here we rewrite this problem in terms of local parameters $\boldsymbol{a}$ and S.

Lemma

For any $\boldsymbol{x}$ and any $\mathbb{Z}$, it holds

$$\mathscr{K}(\mathbb{P}_{\boldsymbol{x},\mathbb{Z}} \,\|\, \mathbb{P}_f) = \mathtt{C} + \frac{1}{2}\log\det(\mathbb{Z}^{-1}) - \frac{p}{2} - \mathbb{E} f(\boldsymbol{x} + \boldsymbol{\gamma}_{\mathbb{Z}}) \,.$$

with $\mathtt{C}$ depending on f and p only.

With $\mathtt{C}_f \stackrel{\mathrm{def}}{=} \log \int \mathrm{e}^{f(\bar{\boldsymbol{x}}+\boldsymbol{u})}\, d\boldsymbol{u}$ and $\mathtt{C}_p = (2\pi)^{-p/2}$, for any $\boldsymbol{u} \in I\!R^p$

$$\frac{dI\!P_f}{d\boldsymbol{u}}(\boldsymbol{x}+\boldsymbol{u}) = \mathrm{e}^{-\mathtt{C}_f}\, \mathrm{e}^{f(\boldsymbol{x}+\boldsymbol{u})},$$

$$\frac{dI\!P_{\boldsymbol{x},\mathbb{Z}}}{d\boldsymbol{u}}(\boldsymbol{x}+\boldsymbol{u}) = \mathtt{C}_p\, \det(\mathbb{Z}^{1/2})\, \mathrm{e}^{-\|\mathbb{Z}^{1/2}\boldsymbol{u}\|^2/2}.$$

This yields with $\boldsymbol{\gamma}_{\mathbb{Z}} \sim \mathcal{N}(0, \mathbb{Z}^{-1})$ and $\boldsymbol{\gamma} \sim \mathcal{N}(0, I\!\!I_p)$

$$I\!E_{\boldsymbol{x},\mathbb{Z}} \log \frac{dI\!P_{\boldsymbol{x},\mathbb{Z}}}{dI\!P_f}$$

$$= \mathtt{C}_f + \log \mathtt{C}_p - I\!E f(\boldsymbol{x}+\boldsymbol{\gamma}_{\mathbb{Z}}) - \frac{1}{2} I\!E\|\boldsymbol{\gamma}\|^2 - \frac{1}{2} \log\det(\mathbb{Z}^{-1}),$$

and the result follows in view of $I\!E\|\boldsymbol{\gamma}\|^2 = p$.

With $\mathbb{F} = -\nabla^2 f(\bar{x})$, represent $\mathbb{Z}$ in the form

$$\mathbb{Z}^{-1/2} = \mathbb{F}^{-1/2} + S \quad \text{or} \quad \mathbb{F}^{1/4}\, \mathbb{Z}^{-1/2}\, \mathbb{F}^{1/4} = I\!\!I_p + \mathbb{F}^{1/4}\, S\, \mathbb{F}^{1/4}.$$

A vicinity of $\mathbb{F}$ using Kullback-Leibler divergence $\mathscr{K}(I\!\!N_{\bar{x},\mathbb{F}} \,\|\, I\!\!N_{\bar{x},\mathbb{Z}})$.

Lemma

Let $\mathbb{Z}^{-1/2} = \mathbb{F}^{-1/2} + S$ *and* $U = \mathbb{F}^{1/4}\, S\, \mathbb{F}^{1/4}$ *fulfill* $\|U\| \le \nu < 1$. *Then*

$$\begin{aligned} &\mathscr{K}(I\!\!N_{\bar{x},\mathbb{F}} \,\|\, I\!\!N_{\bar{x},\mathbb{Z}}) \\ &\quad = -\log\det(I\!\!I_p + \mathbb{F}^{1/4}\, S\, \mathbb{F}^{1/4}) + \frac{1}{2}\operatorname{tr}\bigl\{\mathbb{F}(\mathbb{F}^{-1/2} + S)^2 - I\!\!I_p\bigr\} \\ &\quad = -\log\det(I\!\!I_p + U) + \operatorname{tr} U + \frac{1}{2}\operatorname{tr}(\mathbb{F} S^2) \ge \frac{1}{2}\operatorname{tr}(\mathbb{F} S^2). \end{aligned} \tag{4}$$

For two Gaussian distributions $\mathbb{N}_{\bar{\boldsymbol{x}},\mathbb{F}}, \mathbb{N}_{\bar{\boldsymbol{x}},\mathbb{Z}}$ with the same mean $\bar{\boldsymbol{x}}$

$$\mathscr{K}(\mathbb{N}_{\bar{\boldsymbol{x}},\mathbb{F}} \,\|\, \mathbb{N}_{\bar{\boldsymbol{x}},\mathbb{Z}}) = \frac{1}{2}\bigl\{-\log\det(\mathbb{F}\,\mathbb{Z}^{-1}) + \operatorname{tr}(\mathbb{F}\mathbb{Z}^{-1} - I\!I_p)\bigr\}$$

$$= -\log\det\bigl\{\mathbb{F}^{1/2}(\mathbb{F}^{-1/2} + S)\bigr\} + \frac{1}{2}\operatorname{tr}\bigl\{\mathbb{F}(\mathbb{F}^{-1/2} + S)^2 - I\!I_p\bigr\}$$

$$= -\log\det(I\!I_p + U) + \frac{1}{2}\operatorname{tr}(\mathbb{F}S^2 + 2\mathbb{F}^{1/2}S)$$

and (4) follows by $x - \log(1 + x) \ge 0$ for any $x > -1$.

Consider symmetric matrices $S \in \mathfrak{M}_p$ such that for some $\nu < 1$

$$\|\mathbb{F}^{1/4} S \, \mathbb{F}^{1/4}\| \le \nu \, . \tag{5}$$

Lemma

With $\gamma \sim \mathcal{N}(0, I\!I_p)$, $\boldsymbol{a} \in I\!R^p$, *and* $S \in \mathfrak{M}_p$ *satisfying* (5), *define*

$$H(\boldsymbol{a}, S) \stackrel{\text{def}}{=} -\log\det(\mathbb{F}^{-1/2} + S) - I\!E f(\bar{\boldsymbol{x}} + \boldsymbol{a} + (\mathbb{F}^{-1/2} + S)\boldsymbol{\gamma}),$$

$$(\widehat{\boldsymbol{a}}, \widehat{S}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{(\boldsymbol{a}, S)} H(\boldsymbol{a}, S) \, .$$

Then the VI problem leads to minimization of the function $H(\boldsymbol{a}, S)$ *:*

$$(\widehat{\boldsymbol{x}}, \widehat{\mathbb{Z}}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{(\boldsymbol{x}, \mathbb{Z})} \mathscr{K}(I\!P_{\boldsymbol{x}, \mathbb{Z}} \, \| \, I\!P_f) = \bigl(\bar{\boldsymbol{x}} + \widehat{\boldsymbol{a}}, (\mathbb{F}^{-1/2} + \widehat{S})^{-2}\bigr).$$

For $X \sim I\!P_f \propto \mathrm{e}^{f(\boldsymbol{x})}$, consider

$$\bar{\boldsymbol{x}} = I\!E_f X, \quad \Sigma = \operatorname{Var}(X), \quad \mathbb{F} = -\nabla^2 f(\bar{\boldsymbol{x}}).$$

Consider

$$H(\boldsymbol{a}, S) \stackrel{\text{def}}{=} -\log\det(\mathbb{F}^{-1/2} + S) - I\!E f(\bar{\boldsymbol{x}} + \boldsymbol{a} + (\mathbb{F}^{-1/2} + S)\boldsymbol{\gamma}),$$

$$(\widehat{\boldsymbol{a}}, \widehat{S}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{(\boldsymbol{a},S)} H(\boldsymbol{a}, S).$$

A guess $(\boldsymbol{a}, S) = (0, 0)$. How far from the solution $(\widehat{\boldsymbol{a}}, \widehat{S})$?

Technical issue: anisotropic smoothness in $\boldsymbol{a}$ and S directions.

Fix $\mathbb{Z}^{-1/2} = \mathbb{F}^{-1/2} + S$ and optimize w.r.t. $\boldsymbol{a}$.

For $\mathbb{Z}^{-1/2} = \mathbb{F}^{-1/2} + S$ fixed, consider $H(\boldsymbol{a}) = H(\boldsymbol{a}, S)$

$$\widehat{\boldsymbol{a}} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{a}} H(\boldsymbol{a}) = \operatorname*{argmax}_{\boldsymbol{a}} \mathbb{E} f(\bar{\boldsymbol{x}} + \boldsymbol{a} + \mathbb{Z}^{-1/2}\boldsymbol{\gamma}) \, .$$

Main step: compute $\boldsymbol{A} = \nabla H(0)$ and $\mathscr{F} = -\nabla^2 H(0)$.

A guess: $\mathscr{F} \approx \mathbb{F} = -\nabla^2 f(\bar{\boldsymbol{x}})$, $\boldsymbol{A} \approx 0$ up to fourth order.

Fix $\boldsymbol{a}$ and consider

$$h(t) = -\mathbb{E} f(\bar{\boldsymbol{x}} + t\boldsymbol{a} + \mathbb{Z}^{-1/2}\boldsymbol{\gamma}).$$

Lemma

The function $h(t) = H(t\boldsymbol{a})$ *is strongly convex and satisfies*

$$h''(t) = -\left\langle \mathbb{E}\nabla^2 f(\bar{\boldsymbol{x}} + t\boldsymbol{a} + \mathbb{Z}^{-1/2}\boldsymbol{\gamma}), \boldsymbol{a}^{\otimes 2} \right\rangle .$$

Concavity of $f(\cdot)$ implies convexity of h.

Lemma

It holds with $\mathbb{F} = -\nabla^2 f(\bar{\boldsymbol{x}})$

$$h''(0) = -\mathbb{E} \left\langle \nabla^2 f(\bar{\boldsymbol{x}} + \boldsymbol{\gamma}_{\mathbb{Z}}), \boldsymbol{a}^{\otimes 2} \right\rangle ,$$

and with $\mathrm{p} = \operatorname{tr}(\mathbb{D}\, \mathbb{F}^{-1} \mathbb{D})$ *and* $\alpha = \|\mathbb{D}\, \mathbb{F}^{-1} \mathbb{D}\|$

$$\left| h''(0) - \boldsymbol{a}^{\top} \mathbb{F} \boldsymbol{a} \right| \le \frac{\tau_4 (\mathrm{p} + 2\alpha)}{2} \, \|\mathbb{D} \boldsymbol{a}\|^2 . \qquad (6)$$

It holds

$$-\langle \nabla^2 f(\bar{\boldsymbol{x}}), \boldsymbol{a}^{\otimes 2} \rangle = \boldsymbol{a}^\top \mathbb{F} \boldsymbol{a} .$$

For any $\boldsymbol{u} \in \mathbb{R}^p$,

$$\left| -\langle \nabla^2 f(\bar{\boldsymbol{x}} + \boldsymbol{\gamma}_{\mathbb{Z}}), \boldsymbol{u}^{\otimes 2} \rangle + \langle \nabla^2 f(\bar{\boldsymbol{x}}), \boldsymbol{u}^{\otimes 2} \rangle + \langle \nabla^3 f(\bar{\boldsymbol{x}}), \boldsymbol{\gamma}_{\mathbb{Z}} \otimes \boldsymbol{u}^{\otimes 2} \rangle \right| \\ \leq \frac{1}{2} \tau_4 \, \|\mathbb{D} \boldsymbol{\gamma}_{\mathbb{Z}}\|^2 \, \|\mathbb{D} \boldsymbol{u}\|^2 .$$

With $\mathrm{p} = \mathrm{tr}(\mathbb{D}^2 \mathbb{F}^{-1})$

$$\mathbb{E} \|\mathbb{D} \boldsymbol{\gamma}_{\mathbb{Z}}\|^2 = \mathrm{p} .$$

Further, $\mathbb{E} \langle \nabla^3 f(\bar{\boldsymbol{x}}), \boldsymbol{\gamma}_{\mathbb{Z}} \otimes \boldsymbol{a}^{\otimes 2} \rangle = 0$ and (6) follows.

Define for any direction $\boldsymbol{a}$

$$h(t) = -\mathbb{E} f(\bar{\boldsymbol{x}} + t\boldsymbol{a} + \mathbb{Z}^{-1/2}\boldsymbol{\gamma}).$$

Lemma

It holds with $\mathrm{p} = \operatorname{tr}(\mathbb{D}\, \mathbb{F}^{-1}\mathbb{D})$, $\alpha = \|\mathbb{D}\, \mathbb{F}^{-1}\mathbb{D}\|$,

$$\bigl|h'(0)\bigr| \le \frac{\tau_4\,(\mathrm{p}+\alpha)^{3/2}}{6}\,\|\mathbb{D}\boldsymbol{a}\| + \frac{\diamondsuit_{4,1}}{1-\diamondsuit}\,\|\mathbb{D}\boldsymbol{a}\|\,.$$

With $\gamma_{\mathbb{Z}} = \mathbb{Z}^{-1/2}\gamma$, Taylor expansion of $\nabla f(\bar{\boldsymbol{x}} + \boldsymbol{\gamma}_{\mathbb{Z}})$ yields for any $\boldsymbol{u} \in I\!\!R^p$

$$\Bigl|\langle \nabla f(\bar{\boldsymbol{x}} + \boldsymbol{\gamma}_{\mathbb{Z}}), \boldsymbol{u}\rangle - \langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{u}\rangle - \langle \nabla^2 f(\bar{\boldsymbol{x}}), \boldsymbol{\gamma}_{\mathbb{Z}} \otimes \boldsymbol{u}\rangle - \frac{1}{2}\langle \nabla^3 f(\bar{\boldsymbol{x}}), \boldsymbol{\gamma}_{\mathbb{F}} \otimes \boldsymbol{\gamma}_{\mathbb{Z}} \otimes \boldsymbol{u}\rangle\Bigr| \le \frac{1}{6}\, \tau_4 \, \|\mathbb{D}\boldsymbol{\gamma}_{\mathbb{Z}}\|^3 \, \|\mathbb{D}\boldsymbol{u}\| \, . \tag{7}$$

Also by Laplace approximation

$$\Bigl|\nabla f(\bar{\boldsymbol{x}}), \boldsymbol{a}\rangle - \frac{1}{2} I\!\!E \langle \nabla^3 f(\bar{\boldsymbol{x}}), \boldsymbol{\gamma}_{\mathbb{F}} \otimes \boldsymbol{\gamma}_{\mathbb{F}} \otimes \boldsymbol{a}\rangle\Bigr| \le \frac{\diamondsuit_{4,1}}{1 - \diamondsuit} \, \|\mathbb{D}\boldsymbol{a}\| \, .$$

Now we apply (7) with $\boldsymbol{u} = \boldsymbol{a}$ and $I\!\!E \|\mathbb{D}\boldsymbol{\gamma}_{\mathbb{F}}\|^3 \le (\mathrm{p} + \alpha)^{3/2}$. The use of $I\!\!E \langle \nabla^2 f(\bar{\boldsymbol{x}}), \boldsymbol{\gamma}_{\mathbb{F}} \otimes \boldsymbol{a}\rangle = 0$ yields

$$\Bigl| I\!\!E \, \langle \nabla f(\bar{\boldsymbol{x}} + \mathbb{Z}^{-1/2}\gamma), \boldsymbol{a}\rangle\Bigr| \le \frac{\tau_4 \, (\mathrm{p} + \alpha)^{3/2}}{6} \, \|\mathbb{D}\boldsymbol{a}\| + \frac{\diamondsuit_{4,1}}{1 - \diamondsuit} \, \|\mathbb{D}\boldsymbol{a}\|.$$

Theorem (3-bound)

$$\left\| \mathbb{F}^{1/2}\widehat{\boldsymbol{a}} - \mathbb{F}^{-1/2}\boldsymbol{A} \right\| \leq \; \tau_3 \|\mathbb{F}^{-1/2}\boldsymbol{A}\|^3$$

Theorem (4-bound)

$$\left\| \mathbb{F}^{1/2}\widehat{\boldsymbol{a}} - \mathbb{F}^{-1/2}\boldsymbol{A} - \mathbb{F}^{-1/2}\nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A}) \right\| \leq \; \mathtt{C}\,(\tau_3^2 + \tau_4)\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^3 .$$

References

Alquier, P. and Ridgway, J. (2020).
Concentration of tempered posteriors and of their variational approximations.
The Annals of Statistics, 48(3):1475 – 1497.

Banach, S. (1938).
Über homogene Polynome in (L^2).
Studia Mathematica, 7(1):36–44.

Challis, E. and Barber, D. (2013).
Gaussian kullback-leibler approximate inference.
Journal of Machine Learning Research, 14(68):2239–2286.

David M. Blei, A. K. and McAuliffe, J. D. (2017).
Variational inference: A review for statisticians.
Journal of the American Statistical Association, 112(518):859–877.

Gratton, S., Lawless, A. S., and Nichols, N. K. (2007).
Approximate gauss-newton methods for nonlinear least squares problems.
SIAM J. Optim., 18:106–132.

Han, W. and Yang, Y. (2019).
Statistical inference in mean-field variational bayes.
https://arxiv.org/abs/1911.01525.

Katsevich, A. and Rigollet, P. (2023).
On the approximation accuracy of gaussian variational inference.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. (2023).
Variational inference via wasserstein gradient flows.
https://arxiv.org/abs/2205.15902.

Spokoiny, V. (2024a).
Deviation bounds for the norm of a random vector under exponential moment conditions with applications.
https://arxiv.org/abs/2309.02302v1.

Spokoiny, V. (2024b).
Sharp deviation bounds and concentration phenomenon for the squared norm of a sub-gaussian vector.
https://arxiv.org/abs/2305.07885.

Wang, Y. and Blei, D. M. (2019).
Frequentist consistency of variational bayes.
Journal of the American Statistical Association, 114(527):1147–1161.

Zhang, F. and Gao, C. (2020).
Convergence rates of variational posterior distributions.
The Annals of Statistics, 48(4):2180 – 2207.