Weierstraß-Institut für
Angewandte Analysis und Stochastik

Leibniz
Gemeinschaft

Gaussian Variational Inference

Vladimir Spokoiny ,
WIAS, HU Berlin

10. Oktober 2024

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. $\mathbb{N}_{\boldsymbol{x},\mathbb{Z}} \stackrel{\text{def}}{=} \mathcal{N}(\boldsymbol{x}, \mathbb{Z}^{-1})$.

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

Natural candidates:

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

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).

- [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 $\mathscr{K}\left(\mathbb{N}_{\boldsymbol{x},\mathbb{Z}} \,\|\, \mathbb{P}_f\right)$ 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

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

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

With $\mathtt{C}_f \stackrel{\text{def}}{=} \log \int \mathrm{e}^{f(\bar{\boldsymbol{x}}+\boldsymbol{u})} \, d\boldsymbol{u}$ 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 $\mathscr{K}\,(\mathbb{N}_{\bar{\boldsymbol{x}},\mathbb{F}} \,\|\, \mathbb{N}_{\bar{\boldsymbol{x}},\mathbb{Z}})$.

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*

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

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

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

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

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

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

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

$$\|\mathbb{F}^{1/4} S \, \mathbb{F}^{1/4}\| \leq \nu \, . \tag{2}$$

Lemma

With $\boldsymbol{\gamma} \sim \mathcal{N}(0, I\!\!I_p)$, $\boldsymbol{a} \in I\!\!R^p$ *, and* $S \in \mathfrak{M}_p$ *satisfying* (2)*, 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 $H(\boldsymbol{a}, S)$ *:*

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

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

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

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

$$g(\boldsymbol{v}) - g(\boldsymbol{v}^*) = \langle \boldsymbol{v} - \boldsymbol{v}^*, \boldsymbol{A} \rangle + f(\boldsymbol{v}) - f(\boldsymbol{v}^*). \tag{3}$$

Define

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

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

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 in $\boldsymbol{\upsilon}$ and $\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E} L(\boldsymbol{\upsilon})$ is linear in $\boldsymbol{\upsilon}$:

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

Let $h(\cdot)$ be concave and

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

Consider

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

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

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

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 $\boldsymbol{A} = \nabla f(\boldsymbol{v}^\circ)$.

Lemma

Let $f(\boldsymbol{\upsilon})$ *be quadratic with* $\nabla^2 f(\boldsymbol{\upsilon}) \equiv -\mathbb{F}$ *. If* $g(\boldsymbol{\upsilon})$ *satisfy* (3)*, then*

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

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

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

Further, (3) and $\nabla f(\boldsymbol{\upsilon}^{*}) = 0$ yield $\nabla g(\boldsymbol{\upsilon}^{*}) = \boldsymbol{A}$. Together with $\nabla g(\boldsymbol{\upsilon}^{\circ}) = 0$, this implies $\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*} = \mathbb{F}^{-1}\boldsymbol{A}$.

Taylor expansion of g at $\boldsymbol{\upsilon}^{\circ}$ yields by $\nabla g(\boldsymbol{\upsilon}^{\circ}) = 0$

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

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

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

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

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

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

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

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

Proposition

Under $(\mathcal{T}_{\mathbf{3}}^{*})$

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

and

$$\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*} - \mathbb{F}^{-1}\boldsymbol{A})\| \le \frac{3\tau_3}{4}\|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 . \tag{5}$$

Implies Newton – Kantorovich – Nemirovskii-Nesterov results about quadratic convergence of second order methods.

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

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

Further,

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

As $\boldsymbol{\upsilon}^{\circ} \in \mathcal{A}(\mathtt{r})$ and it maximizes $g(\boldsymbol{\upsilon})$, we derive by (6) and Lemma 5

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

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

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

Now (4) 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 (5) 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 **??** with $\boldsymbol{u} = \mathbb{F}^{-1}\boldsymbol{A}$

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

Further, by (3)

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

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

$$\|\mathbb{F}^{-1/2}\{\nabla g(\boldsymbol{v}^* + \mathbb{F}^{-1}\boldsymbol{A}) - \nabla g(\boldsymbol{v}^\circ)\}\| \le \frac{\tau_3}{2} \|\mathbb{F}^{-1/2}\boldsymbol{A}\|^2 . \tag{7}$$

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 (3) $\nabla^2 g(\boldsymbol{\upsilon}) = \nabla^2 f(\boldsymbol{\upsilon})$ 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 (7) 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 (5) follows.

Lemma

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

$$\max_{\|\boldsymbol{u}\| \le \mathtt{r}} \left(\frac{\tau}{3} \|\boldsymbol{u}\|^3 - \|\boldsymbol{u} - \boldsymbol{\xi}\|^2 \right) \le \frac{\tau}{2} \|\boldsymbol{\xi}\|^3 , \tag{8}$$

$$\min_{\|\boldsymbol{u}\| \le \mathtt{r}} \left(\frac{\tau}{3} \|\boldsymbol{u}\|^3 + \|\boldsymbol{u} - \boldsymbol{\xi}\|^2 \right) \le \frac{\tau}{3} \|\boldsymbol{\xi}\|^3 . \tag{9}$$

Any maximizer $\boldsymbol{u}$ of the left hand-side of (8) 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:

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

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,

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

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

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 $y \geq 1$. As $\phi(1/3) \leq 3/2$, (8) follows.

The proof of (9) 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 $a \in [0, 1/3]$

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

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

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

Assume the conditions of Proposition 1 and $(\mathcal{T}_4^*)$ *. Then* $\boldsymbol{\upsilon}^{\circ} = \operatorname{argmax}_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon})$ *satisfies* $\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^{\circ} - \boldsymbol{\upsilon}^{*})\| \le \mathtt{r}$ *. With*

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

it holds

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

Also with $\boldsymbol{\xi} = \mathbb{F}^{-1/2}\boldsymbol{A}$

$$\Bigl| g(\boldsymbol{\upsilon}^{\circ}) - g(\boldsymbol{\upsilon}^{*}) - \frac{\|\boldsymbol{\xi}\|^2}{2} - \mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A}) \Bigr| \le \frac{\tau_4 + 7\tau_3^2}{16} \|\boldsymbol{\xi}\|^4 + \frac{(\tau_4 + 3\tau_3^2)^2}{5} \|\boldsymbol{\xi}\|^6 .$$

W.l.o.g. assume $\boldsymbol{v}^* = 0$. It holds by $(\boldsymbol{\mathcal{T}_3^*})$

$$\|\mathbb{F}^{1/2}\boldsymbol{a} - \boldsymbol{\xi}\| = \|\mathbb{F}^{-1/2}\nabla\mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\|$$

$$= \sup_{\|\boldsymbol{u}\|=1} 3\bigl|\langle\mathcal{T}, \mathbb{F}^{-1}\boldsymbol{A} \otimes \mathbb{F}^{-1}\boldsymbol{A} \otimes \mathbb{F}^{-1/2}\boldsymbol{u}\rangle\bigr| \le \frac{\tau_3}{2}\|\boldsymbol{\xi}\|^2 \tag{11}$$

yielding by $\|\boldsymbol{\xi}\| \le \nu\, \mathtt{r}$

$$\|\mathbb{F}^{1/2}\boldsymbol{a}\| \le \Bigl(1 + \frac{\tau_3\, \nu\, \mathtt{r}}{2}\Bigr)\|\boldsymbol{\xi}\|\,. \tag{12}$$

Similarly for any $\boldsymbol{v}$

$$\|\mathbb{F}^{-1/2}\, \nabla^2\mathcal{T}(\mathbb{F}^{-1/2}\boldsymbol{v})\, \mathbb{F}^{-1/2}\| \le \tau_3\|\boldsymbol{v}\|\,.$$

Proof

Furthermore, the tensor $\nabla^2 \mathcal{T}(\boldsymbol{u})$ is linear in $\boldsymbol{u}$ and hence,

$$\sup_{t\in[0,1]} \|\mathbb{F}^{-1/2}\, \nabla^2 \mathcal{T}(t\boldsymbol{a} + (1-t)\mathbb{F}^{-1}\boldsymbol{A})\, \mathbb{F}^{-1/2}\|$$

$$= \max\{\|\mathbb{F}^{-1/2}\, \nabla^2 \mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\, \mathbb{F}^{-1/2}\|, \|\mathbb{F}^{-1}\nabla^2 \mathcal{T}(\boldsymbol{a})\|\}$$

$$\le \tau_3 \max\{\|\boldsymbol{\xi}\|, \|\mathbb{F}^{1/2}\boldsymbol{a}\|\}\,.$$

Later we assume $\|\mathbb{F}^{1/2}\boldsymbol{a}\| \ge \|\mathbb{F}^{-1}\boldsymbol{A}\|$ in view of (12). This and (11) yield

$$\|\mathbb{F}^{-1/2}\nabla \mathcal{T}(\boldsymbol{a}) - \mathbb{F}^{-1/2}\nabla \mathcal{T}(\mathbb{F}^{-1}\boldsymbol{A})\|$$

$$\le \sup_{t\in[0,1]} \|\mathbb{F}^{-1/2}\, \nabla^2 \mathcal{T}(t\boldsymbol{a} + (1-t)\boldsymbol{A})\, \mathbb{F}^{-1/2}\|\ \|\mathbb{F}^{1/2}\boldsymbol{a} - \boldsymbol{\xi})\| \le \frac{\tau_3^2}{2}\, \|\mathbb{F}^{1/2}\boldsymbol{a}\|^3$$

Further, in view of $\nabla \mathcal{T}(\boldsymbol{a}) = \frac{1}{2}\langle \nabla^3 f(\boldsymbol{v}^*), \boldsymbol{a} \otimes \boldsymbol{a}\rangle$

$$\left\|\mathbb{F}^{-1/2}\{\nabla f(\boldsymbol{a}) + \mathbb{F}\boldsymbol{a} - \nabla \mathcal{T}(\boldsymbol{a})\}\right\| \le \frac{\tau_4}{6}\|\mathbb{F}^{1/2}\boldsymbol{a}\|^3\,.$$

Now we can bound the norm of $\mathbb{F}^{-1/2}\nabla g(\boldsymbol{a})$. In view of (3), it holds

$$
\begin{aligned}
\left\|\mathbb{F}^{-1/2}\nabla g(\boldsymbol{a})\right\| &= \left\|\mathbb{F}^{-1/2}\{\nabla g(\boldsymbol{a}) + \mathbb{F}\boldsymbol{a} - \nabla\mathcal{T}(\boldsymbol{A}) - \boldsymbol{A}\}\right\| \\
&\le \left\|\mathbb{F}^{-1/2}\{\nabla f(\boldsymbol{a}) + \mathbb{F}\boldsymbol{a} - \nabla\mathcal{T}(\boldsymbol{a})\}\right\| + \left\|\mathbb{F}^{-1/2}\{\nabla\mathcal{T}(\boldsymbol{a}) - \nabla\mathcal{T}(\boldsymbol{A})\}\right\| \\
&\le \frac{\tau_4 + 3\tau_3^2}{6}\|\mathbb{F}^{1/2}\boldsymbol{a}\|^3.
\end{aligned}
$$

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

$$
\|\mathbb{F}^{-1/2}\{\nabla g(\boldsymbol{a}) - \nabla g(\boldsymbol{\upsilon}^{\circ})\}\| \le \frac{\tau_4 + 3\tau_3^2}{6}\|\mathbb{F}^{1/2}\boldsymbol{a}\|^3 . \tag{13}
$$

Furthermore, with $\varDelta = \boldsymbol{a} - \boldsymbol{\upsilon}^{\circ}$

$$
\mathbb{F}^{-1/2}\{\nabla g(\boldsymbol{a}) - \nabla g(\boldsymbol{\upsilon}^{\circ})\} = \left(\int_0^1 \mathbb{F}^{-1/2}\,\nabla^2 g(\boldsymbol{\upsilon}^{\circ} + t\varDelta)\,\mathbb{F}^{-1/2}\,dt\right)\mathbb{F}^{1/2}\varDelta\,.
$$

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

$$\|\mathbb{F}^{-1/2}\{\nabla g(\boldsymbol{a}) - \nabla g(\boldsymbol{\upsilon}^\circ)\}\| \ge (1 - \tau_3\, \mathtt{r})\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^\circ - \boldsymbol{a})\| .$$

This, (12), and (13) yield in view of $\tau_3\, \mathtt{r} \le 1/3$ and $\nu = 2/3$

$$\|\mathbb{F}^{1/2}(\boldsymbol{\upsilon}^\circ - \boldsymbol{a})\| \le \frac{\tau_4 + 3\tau_3^2}{6(1 - \tau_3\, \mathtt{r})}\|\mathbb{F}^{1/2}\boldsymbol{a}\|^3 \le \frac{\tau_4 + 3\tau_3^2}{3}\|\boldsymbol{\xi}\|^3 . \qquad (14)$$

It remains to bound $g(\boldsymbol{\upsilon}^{\circ}) - g(0)$. By (11)

$$\frac{1}{2}\|\boldsymbol{\xi}\|^2 - \langle \boldsymbol{A}, \boldsymbol{a} \rangle + \frac{1}{2}\|\mathbb{F}^{1/2}\boldsymbol{a}\|^2 = \frac{1}{2}\|\mathbb{F}^{1/2}\boldsymbol{a} - \boldsymbol{\xi}\|^2 \leq \frac{\tau_3^2}{8}\|\boldsymbol{\xi}\|^4 .$$

First consider $g(\boldsymbol{a}) - g(0)$. One more use of $(\boldsymbol{\mathcal{T}_4^*})$ yields with $\boldsymbol{\upsilon}^* = 0$ and $-\nabla^2 f(0) = \mathbb{F}$

$$\begin{aligned}
&\Bigl| g(\boldsymbol{a}) - g(0) - \frac{1}{2}\|\boldsymbol{\xi}\|^2 - \mathcal{T}(\boldsymbol{a}) \Bigr| \\
&= \Bigl| f(\boldsymbol{a}) - f(0) + \langle \boldsymbol{A}, \boldsymbol{a} \rangle - \frac{1}{2}\|\boldsymbol{\xi}\|^2 - \mathcal{T}(\boldsymbol{a}) \Bigr| \\
&\leq \Bigl| f(\boldsymbol{a}) - f(0) + \frac{1}{2}\|\mathbb{F}^{1/2}\boldsymbol{a}\|^2 - \mathcal{T}(\boldsymbol{a}) \Bigr| + \frac{\tau_3^2}{8}\|\boldsymbol{\xi}\|^4 \\
&\leq \frac{\tau_4}{24}\|\mathbb{F}^{1/2}\boldsymbol{a}\|^4 + \frac{\tau_3^2}{8}\|\boldsymbol{\xi}\|^4 \leq \frac{\tau_4 + 2\tau_3^2}{16}\|\boldsymbol{\xi}\|^4 .
\end{aligned}$$

Also by $\nabla g(\boldsymbol{v}^{\circ}) = 0$ and (14), it holds for some $\boldsymbol{v} \in [\boldsymbol{a}, \boldsymbol{v}^{\circ}]$ as in (14)

$$\begin{aligned} \big| g(\boldsymbol{a}) - g(\boldsymbol{v}^{\circ}) \big| &\leq \frac{1}{2} \| \mathbb{F}^{-1/2} \, \nabla^2 g(\boldsymbol{v}) \, \mathbb{F}^{-1/2} \| \; \| \mathbb{F}^{1/2} (\boldsymbol{a} - \boldsymbol{v}^{\circ}) \|^2 \\ &\leq \frac{(\tau_4 + 3\tau_3^2)^2}{72 (1 - \tau_3 \, \mathtt{r})^3} \, \| \mathbb{F}^{1/2} \boldsymbol{a} \|^6 < \frac{(\tau_4 + 3\tau_3^2)^2}{5} \, \| \boldsymbol{\xi} \|^6 \, , \end{aligned}$$

Moreover, similarly to (11)

$$\begin{aligned} \big| \mathcal{T}(\boldsymbol{a}) - \mathcal{T}(\mathbb{F}^{-1} \boldsymbol{A}) \big| &\leq \sup_{t \in [0,1]} \big\| \mathbb{F}^{-1/2} \nabla \mathcal{T}(t \mathbb{F}^{-1} \boldsymbol{A} + (1 - t) \boldsymbol{a}) \big\| \; \big\| \mathbb{F}^{1/2} \boldsymbol{a} - \boldsymbol{\xi} \big| \\ &\leq \frac{\tau_3^2}{4} \| \mathbb{F}^{1/2} \boldsymbol{a} \|^2 \, \| \boldsymbol{\xi} \|^2 \leq \frac{5 \tau_3^2}{16} \| \boldsymbol{\xi} \|^4 \, . \end{aligned}$$

Summing up the obtained bounds yields (10).

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 $f(\boldsymbol{v})$ *be quadratic with* $\mathbb{F} \equiv -\nabla^2 f(\boldsymbol{v})$ *and* $\boldsymbol{S}_G \equiv G^2 \boldsymbol{v}^*$ *. Then it holds with* $\mathbb{F}_G = \mathbb{F} + G^2$

$$\boldsymbol{v}^* - \boldsymbol{v}_G^* = \quad \mathbb{F}_G^{-1} \boldsymbol{S}_G \quad = -\mathbb{F}_G^{-1} G^2 \boldsymbol{v}^*,$$

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

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

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

Further, $\nabla f(\boldsymbol{v}^*) = 0$ yielding $\nabla f_G(\boldsymbol{v}^*) = \boldsymbol{S}_G = G^2 \boldsymbol{v}^*$. Together with $\nabla f_G(\boldsymbol{v}_G^*) = 0$, this implies $\boldsymbol{v}^* - \boldsymbol{v}_G^* = \mathbb{F}_G^{-1} \boldsymbol{S}_G$. The Taylor expansion of f_G at $\boldsymbol{v}_G^*$ yields

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

and the assertion follows.

Proposition

Let f *be concave and* $\boldsymbol{\upsilon}^* = \operatorname{argmax}_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon})$ *. Define* $\boldsymbol{S}_G = G^2 \boldsymbol{\upsilon}^*$ *,*

$$\mathbb{F}_G = -\nabla^2 f(\boldsymbol{\upsilon}^*) + G^2 , \quad \mathsf{b}_G = \| \mathbb{F}_G^{-1/2} G^2 \boldsymbol{\upsilon}^* \| = \| \mathbb{F}_G^{-1/2} \boldsymbol{S}_G \| .$$

With $\nu = 2/3$ *, assume* $(\mathcal{T}_3^*)$ *for* $\mathtt{r} = \nu^{-1} \mathsf{b}_G$ *and* $\mathbb{D}^2 \le \mathbb{F}_G$ *. Then* $\| \mathbb{F}_G^{1/2} (\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*) \| \le \nu^{-1} \, \mathsf{b}_G$ *or, equivalently,*

$$\boldsymbol{\upsilon}_G^* \in \mathcal{A}_G \stackrel{\text{def}}{=} \{ \boldsymbol{\upsilon} \colon \| \mathbb{F}_G^{1/2} (\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*) \| \le \nu^{-1} \, \mathsf{b}_G \}. \tag{15}$$

Moreover,

$$\Big\| \mathbb{F}_G^{1/2} (\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^*) - \mathbb{F}_G^{-1/2} \boldsymbol{S}_G \Big\|^2 \le \tau_3 \, \mathsf{b}_G^3 ,$$

$$\Big| 2 f_G(\boldsymbol{\upsilon}_G^*) - 2 f_G(\boldsymbol{\upsilon}^*) - \mathsf{b}_G^2 \Big| \le \tau_3 \, \mathsf{b}_G^3 .$$

With $\boldsymbol{S}_G = G^2\boldsymbol{v}^*$, define $g_G(\boldsymbol{v})$ by

$$g_G(\boldsymbol{v}) - g_G(\boldsymbol{v}_G^*) = f_G(\boldsymbol{v}) - f_G(\boldsymbol{v}_G^*) + \langle \boldsymbol{S}_G, \boldsymbol{v} - \boldsymbol{v}_G^* \rangle. \quad (16)$$

The function f_G is concave, the same holds for g_G from (16). Now we show that $\boldsymbol{v}^* = \operatorname{argmax} g_G(\boldsymbol{v})$. It suffices to check that $\nabla g_G(\boldsymbol{v}^*) = 0$. Indeed, by definition, $\nabla f(\boldsymbol{v}^*) = 0$, and hence, $\nabla f_G(\boldsymbol{v}^*) = -G^2\boldsymbol{v}^* + \boldsymbol{S}_G = 0$. Now the results follow from Proposition 1 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 $\boldsymbol{A} = \boldsymbol{S}_G$.

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

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

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

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

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

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* $\nu = 2/3$ *, assume* $(\mathcal{T}_3^*)$ *and* $(\mathcal{T}_4^*)$ *for* $\mathtt{r} = \mathtt{r}_G \stackrel{\text{def}}{=} \nu^{-1} \mathsf{b}_G$ *and* $\mathbb{D}^2 \le \mathbb{F}_G$ *. Then* (15) *holds. Furthermore,* $\|\mathbb{F}_G^{1/2} \boldsymbol{m}_G\| \le \mathtt{r}_G$ *and*

$$\|\mathbb{F}^{1/2} \boldsymbol{m}_G - \mathbb{F}_G^{-1/2} \boldsymbol{S}_G\| \le \frac{\tau_3}{2} \|\mathbb{F}_G^{-1/2} \boldsymbol{S}_G\|^2 \le \frac{\tau_3 \, \nu \, \mathtt{r}_G}{2} \|\mathbb{F}_G^{-1/2} \boldsymbol{S}_G\|,$$

$$\|\mathbb{F}^{1/2} \boldsymbol{m}_G\| \le \Bigl(1 + \frac{\tau_3 \, \nu \, \mathtt{r}_G}{2}\Bigr) \|\mathbb{F}_G^{-1/2} \boldsymbol{S}_G\| \,,$$

$$\|\mathbb{F}_G^{1/2} (\boldsymbol{\upsilon}^* - \boldsymbol{\upsilon}_G^* - \boldsymbol{m}_G)\| \le \frac{\tau_4 + 3\tau_3^2}{3} \|\mathbb{F}_G^{-1/2} \boldsymbol{S}_G\|^3 \,.$$

Also

$$\Bigl| f_G(\boldsymbol{\upsilon}_G^*) - f_G(\boldsymbol{\upsilon}^*) - \frac{1}{2} \|\mathbb{F}_G^{-1/2} \boldsymbol{S}_G\|^2 - \mathcal{T}(\boldsymbol{m}_G) \Bigr|$$

$$\le \frac{\tau_4 + 2\tau_3^2}{16} \|\mathbb{F}_G^{-1/2} \boldsymbol{S}_G\|^4 + \frac{(\tau_4 + 3\tau_3^2)^2}{5} \|\mathbb{F}_G^{-1/2} \boldsymbol{S}_G\|^6 \,.$$

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

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

Consider

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

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

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

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

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

$$\left| h''(0) - \boldsymbol{a}^\top \mathbb{F} \boldsymbol{a} \right| \le \frac{\tau_4 (\mathrm{p} + 2\alpha)}{2} \|\mathbb{D} \boldsymbol{a}\|^2 . \tag{17}$$

It holds

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

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

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

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

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

Further, $I\!\!E \langle \nabla^3 f(\bar{\boldsymbol{x}}), \boldsymbol{\gamma}_{\mathbb{Z}} \otimes \boldsymbol{a}^{\otimes 2} \rangle = 0$ and (17) 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 $\mathrm{p} = \operatorname{tr}(\mathbb{D}\,\mathbb{F}^{-1}\mathbb{D})\,,\ \alpha = \|\mathbb{D}\,\mathbb{F}^{-1}\mathbb{D}\|\,,$

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

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

$$\Big| \langle \nabla f(\bar{x} + \gamma_{\mathbb{Z}}), u \rangle - \langle \nabla f(\bar{x}), u \rangle - \langle \nabla^2 f(\bar{x}), \gamma_{\mathbb{Z}} \otimes u \rangle - \frac{1}{2} \langle \nabla^3 f(\bar{x}), \gamma_{\mathbb{F}} \otimes \gamma_{\mathbb{Z}} \otimes u \rangle \Big| \le \frac{1}{6} \tau_4 \, \| \mathbb{D} \gamma_{\mathbb{Z}} \|^3 \, \| \mathbb{D} u \| \,. \tag{18}$$

Also by Laplace approximation

$$\Big| \nabla f(\bar{x}), a \rangle - \frac{1}{2} \mathbb{E} \langle \nabla^3 f(\bar{x}), \gamma_{\mathbb{F}} \otimes \gamma_{\mathbb{F}} \otimes a \rangle \Big| \le \frac{\diamondsuit_{4,1}}{1 - \diamondsuit} \, \| \mathbb{D} a \| \,.$$

Now we apply (18) with $u = a$ and $\mathbb{E} \| \mathbb{D} \gamma_{\mathbb{F}} \|^3 \le (\mathrm{p} + \alpha)^{3/2}$. The use of $\mathbb{E} \langle \nabla^2 f(\bar{x}), \gamma_{\mathbb{F}} \otimes a \rangle = 0$ yields

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

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)

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

Optimization: $f(\boldsymbol{x}) \to \min$.

Sampling: $\boldsymbol{X} \propto \exp\{-f(\boldsymbol{x}) + \log \pi(\boldsymbol{x})\}$ for a sampling density π .

Sampling gradiet free procedure (1 step):

- draw a mini-batch $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from π ;
- compute $w_i = \mathrm{e}^{-f(\boldsymbol{X}_i)}$;
- mini-batch averaging

$$\bar{\boldsymbol{x}}_\pi = \frac{\sum_{i=1}^n \boldsymbol{X}_i w_i}{\sum_{i=1}^n w_i} .$$

Update of π (EnKF, diffusion models, VAE, etc.), loop.

Issues: mini-batch size n , averaging over steps, rate $\|\bar{\boldsymbol{x}}_\pi - \bar{\boldsymbol{x}}\|$.

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.

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.

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.