Supplementary Material for
The Deep Feed-Forward Gaussian Process: An Effective Generalization to Covariance Priors

Melih Kandemir and Fred A. Hamprecht
Heidelberg University, HCI/IWR

Editor: Afshin Rostamizadeh

1. Variational Lower Bound for Classification

Given the variational lower bound $\mathcal{L}_r$ for continuous output, the lower bound for binary output can be calculated by adding the Bernoulli-Probit likelihood $p(\mathbf{t}|\mathbf{y}) = \prod_{n=1}^{N} Bernoulli(t_n|\Phi(y_n))$ to the model, and marginalizing out $\mathbf{y}$. The marginal likelihood for the GP classifier can be bounded by $p(\mathbf{t}|\mathbf{Z},\mathbf{X}) \geq \int \exp(\mathcal{L}_r)p(\mathbf{t}|\mathbf{y})d\mathbf{y}$. After taking this integral, the lower bound becomes

$$\log p(\mathbf{t}|\mathbf{Z},\mathbf{X}) \geq \mathcal{L}_c = \mathcal{L}_r + \sum_{n}^{N} t_n \log \Phi\Big(\frac{\mathbf{m}^T \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbb{E}[\mathbf{K}_{\mathbf{Zb}_n}]}{\sqrt{\beta^{-1}+1}}\Big)$$
$$+ \sum_{n}^{N} t_n \mathbb{I}(t_n = 1) \frac{\beta}{2} (\mathbf{m}^T \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbb{E}[\mathbf{K}_{\mathbf{ZB}}])^2 + \sum_{n}^{N} t_n \mathbb{I}(t_n = -1) \log \sqrt{\frac{2\pi}{\beta}},$$

where $\mathbb{I}(\cdot)$ is the indicator function.

2. Variational Update Rules

For regression, a mean-field update is tractable for $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m},\mathbf{S})$ as follows

$$\mathbf{S} = \Big(\mathbf{K}_{\mathbf{ZZ}}^{-1} + \beta \sum_{n}^{N} tr\{\mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbb{E}_{Q_{AB}}[\mathbf{k}_{\mathbf{ZB}} \mathbf{k}_{\mathbf{ZB}}^T] \mathbf{K}_{\mathbf{ZZ}}^{-1}\}\Big)^{-1},$$
$$\mathbf{m} = \beta \mathbf{S} \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbb{E}_{Q_{AB}}[\mathbf{k}_{\mathbf{ZB}}]\mathbf{y}.$$

However, for classification, this update should be done gradient-based, since $\mathbf{m}$ also appears in the Bernoulli-Probit likelihood in a non-conjugate way. The related gradient equations are

$$\frac{\partial \mathcal{L}_c}{\partial \mathbf{m}} = -\beta {\mathbf{K}_{\mathbf{ZZ}}}^{-1} \mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{ZB}} \mathbf{K}_{\mathbf{ZB}}^T] {\mathbf{K}_{\mathbf{ZZ}}}^{-1}\mathbf{m} - {\mathbf{K}_{\mathbf{ZZ}}}^{-1}\mathbf{m}$$
$$+ \sum_{n=1}^{N} \mathbb{I}(t_n = 1)\beta \Big(\mathbf{m}^T \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{Zb}_n}]\Big) {\mathbf{K}_{\mathbf{ZZ}}}^{-1} \mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{Zb}_n}]$$
$$+ \sum_{n=1}^{N} t_n \frac{\mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{Zb}_n}]}{\sqrt{2\pi}(\beta^{-1}+1)\Phi\left(\frac{\mathbf{m}^T \mathbf{K}_{\mathbf{ZZ}}^{-1} \mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{Zb}_n}]}{\sqrt{\beta^{-1}+1}}\right)},$$

and

$$\frac{\partial \mathcal{L}_c}{\partial \mathbf{S}} = -\frac{1}{2}\mathbf{K_{ZZ}}^{-1} + \frac{1}{2}\mathbf{S}^{-T} - \frac{\beta}{2}\mathbf{K_{ZZ}}^{-1}\mathbb{E}_{Q_{AB}}[\mathbf{K_{ZB}}\mathbf{K}_{\mathbf{ZB}}^T]\mathbf{K_{ZZ}}^{-1}.$$

For both regression and classification, the gradient of the lower bound with respect to $\mathbf{c}_r$ is

$$\begin{aligned}
\frac{\partial \mathcal{L}_r}{\partial \mathbf{c}_r} = &- \mathbf{K}^{-1}_{\mathbf{X}_{ir}\mathbf{X}_{ir}}\mathbf{c}_r + \beta\mathbf{y}^T\frac{\partial \mathbb{E}_{Q_{AB}}[\mathbf{K_{ZB}}]^T}{\partial \mathbf{c}_r}\mathbf{K}^{-1}_{\mathbf{ZZ}}\mathbf{m} \\
&- \frac{\beta}{2}tr\Big\{\mathbf{K}^{-1}_{\mathbf{ZZ}}\frac{\partial \mathbb{E}_{Q_{AB}}[\mathbf{K_{ZB}}\mathbf{K}^T_{\mathbf{ZB}}]}{\partial \mathbf{c}_r}\mathbf{K}^{-1}_{\mathbf{ZZ}}(\mathbf{m}\mathbf{m}^T + \mathbf{S})\Big\} \\
&- \frac{\beta}{2}tr\Big\{\frac{\partial \mathbb{E}_{Q_{AB}}[\mathbf{K_{BB}}]}{\partial \mathbf{c}_r} - \mathbf{K}^{-1}_{\mathbf{ZZ}}\frac{\partial \mathbb{E}_{Q_{AB}}[\mathbf{K_{ZB}}\mathbf{K}^T_{\mathbf{ZB}}]}{\partial \mathbf{c}_r}\Big\}.
\end{aligned}$$

We learn the inducing points by optimizing the lower bound with respect to each entry of $\mathbf{Z}$. The derivative of the lower bound with respect to the inducing point p of DoF r for regression is

$$\begin{aligned}
\frac{\partial \mathcal{L}_r}{\partial z_{pr}} = &\beta\mathbf{y}^T\Big(\frac{\partial \mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{ZB}_n}]}{\partial z_{pr}}^T\mathbf{K}^{-1}_{\mathbf{ZZ}} + \mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{ZB}_n}]^T\frac{\mathbf{K}^{-1}_{\mathbf{ZZ}}}{\partial z_{pr}}\Big)\mathbf{m} \\
&- \frac{\beta}{2}tr\Bigg\{\Big(\frac{\partial \mathbf{K}^{-1}_{\mathbf{ZZ}}}{\partial z_{pr}}\mathbb{E}_{Q_{AB}}[\mathbf{K_{ZB}}\mathbf{K}^T_{\mathbf{ZB}}]\mathbf{K}^{-1}_{\mathbf{ZZ}} + \mathbf{K}^{-1}_{\mathbf{ZZ}}\frac{\partial \mathbb{E}_{Q_{AB}}[\mathbf{K_{ZB}}\mathbf{K}^T_{\mathbf{ZB}}]}{\partial z_{pr}}\mathbf{K}^{-1}_{\mathbf{ZZ}} \\
&\qquad + \mathbf{K}^{-1}_{\mathbf{ZZ}}\mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{Zb}_n}\mathbf{K}^T_{\mathbf{Zb}_n}]\frac{\partial \mathbf{K}^{-1}_{\mathbf{ZZ}}}{\partial z_{pr}}\Big)(\mathbf{m}\mathbf{m}^T + \mathbf{S})\Bigg\} \\
&+ \frac{\beta}{2}tr\Bigg\{\frac{\mathbf{K}^{-1}_{\mathbf{ZZ}}}{\partial z_{pr}}\mathbb{E}_{Q_{AB}}[\mathbf{K_{ZB}}\mathbf{K}^T_{\mathbf{ZB}}] + \mathbf{K}^{-1}_{\mathbf{ZZ}}\frac{\partial \mathbb{E}_{Q_{AB}}[\mathbf{K_{ZB}}\mathbf{K}^T_{\mathbf{ZB}}]}{\partial z_{pr}}\Bigg\} \\
&- \frac{1}{2}tr\Big(\mathbf{K}^{-1}_{\mathbf{ZZ}}\frac{\partial \mathbf{K_{ZZ}}}{\partial z_{pr}} + \frac{\partial \mathbf{K}^{-1}_{\mathbf{ZZ}}}{\partial z_{pr}}\mathbf{S}\Big) - \frac{1}{2}\mathbf{m}^T\frac{\partial \mathbf{K}^{-1}_{\mathbf{ZZ}}}{\partial z_{pr}}\mathbf{m}.
\end{aligned} \tag{1}$$

For classification, this derivative is

$$\frac{\partial \mathcal{L}_c}{\partial z_{pr}} = \frac{\partial \mathcal{L}_r}{\partial z_{pr}} + \sum_{n=1}^{N} t_n\frac{\mathcal{N}(F_n|0,1)}{\Phi(F_n)} + \sum_{n=1}^{N} t_n\frac{\partial F_n}{\partial z_{pr}} + \sum_{n=1}^{N}\mathbb{I}(t_n = 1)\beta F_n\frac{\partial F_n}{\partial z_{pr}}, \tag{2}$$

where $F_n = \frac{\mathbf{m}^T\mathbf{K}^{-1}_{\mathbf{ZZ}}\mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{Zb}_n}]}{\sqrt{\beta^{-1}+1}}$ and

$$\frac{\partial F_n}{\partial z_{pr}} = \mathbf{m}^T\left(\frac{\partial \mathbf{K}^{-1}_{\mathbf{ZZ}}}{\partial z_{pr}}\mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{Zb}_n}] + \mathbf{K}^{-1}_{\mathbf{ZZ}}\frac{\partial \mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{Zb}_n}]}{\partial z_{pr}}\right).$$

Given a Gaussian kernel function $k(\mathbf{z}, \boldsymbol{\mu}) = \exp\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\mathbf{J}^{-1}(\mathbf{z}-\boldsymbol{\mu})\}$, which could be isotropic as in the radial basis function (RBF), diagonal as in Automatic Relevance Determination (ARD), or a full matrix as in metric learning, hyperparameters $\mathbf{J}$ can also be fit by gradient ascent. The derivatives of the variational log-likelihood with respect to the kernel hyperparameters are as in Equations 1 and 2. It suffices to replace all ∂z_{pr}s in these formulas with $\partial \mathbf{J}_{ij}$.

3. RBF Kernel for Random Inputs

For a Radial Basis Function $k(\mathbf{x}, \mathbf{x}') = \exp\{-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^T \mathbf{J}^{-1}(\mathbf{x} - \mathbf{x}')\}$, static input vectors $\mathbf{z}$ and $\mathbf{z}'$, and the random vector $\mathbf{x}$ which follows the multivariate normal distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[k(\mathbf{z}, \mathbf{x})] = |\mathbf{J}^{-1}\boldsymbol{\Sigma} + \mathbf{I}|^{-\frac{1}{2}} \times \exp\Big\{ -\frac{1}{2}\mathbf{z}^T(\mathbf{J} + \boldsymbol{\Sigma})^{-1}\mathbf{z} - \frac{1}{2}\boldsymbol{\mu}^T(\mathbf{J} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\mu} + \mathbf{z}^T\mathbf{J}^{-1}(\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big\},$$

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[k(\mathbf{z}_p, \mathbf{x})k(\mathbf{z}_{p'}, \mathbf{x})] = |2\mathbf{J}^{-1}\boldsymbol{\Sigma} + \mathbf{I}|^{-\frac{1}{2}} \times \exp\Big\{ &-\frac{1}{2}\mathbf{z}_p^T(\mathbf{J}^{-1} - \mathbf{J}^{-1}(2\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\mathbf{J}^{-1})\mathbf{z}_p \\ &-\frac{1}{2}\mathbf{z}_{p'}^T(\mathbf{J}^{-1} - \mathbf{J}^{-1}(2\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\mathbf{J}^{-1})\mathbf{z}_{p'} - \frac{1}{2}\boldsymbol{\mu}^T(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(2\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu} \\ &+ (\mathbf{z}_p + \mathbf{z}_{p'})^T\mathbf{J}^{-1}(2\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{z}_p^T\mathbf{J}^{-1}(2\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\mathbf{J}^{-1}\mathbf{z}_{p'}\Big\}.\end{aligned}$$

The derivatives of the stochastic Gaussian kernel with respect to an inducing point entry z_{pr} are

$$\begin{aligned}\frac{\partial \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[k(\mathbf{z}_p, \boldsymbol{\mu})]}{\partial z_{pr}} =& \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[k(\mathbf{z}_p, \boldsymbol{\mu})]\Big(-\mathbf{z}_p^T\mathbf{J}^{-1} - \mathbf{J}^{-1}(\boldsymbol{\Sigma}^{-1} + \mathbf{J}^{-1})^{-1}\mathbf{J}^{-1} \\ &+ \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^{-1} + \mathbf{J}^{-1})^{-1}\mathbf{J}^{-1}\frac{\partial \mathbf{z}_p}{\partial z_{pr}}\Big),\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[k(\mathbf{z}_p, \boldsymbol{\mu})k(\mathbf{z}_{p'}, \boldsymbol{\mu})]}{\partial z_{pr}} =& \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[k(\mathbf{z}_p, \boldsymbol{\mu})k(\mathbf{z}_{p'}, \boldsymbol{\mu})] \\ &\times \Bigg(-\mathbf{z}_p^T(\mathbf{J}^{-1} - \mathbf{J}^{-1}(2\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\mathbf{J}^{-1})\frac{\partial \mathbf{z}_p}{\partial z_{pr}} \\ &+ \Big(\frac{\partial \mathbf{z}_p}{\partial z_{pr}} + \mathbf{z}_{p'}\Big)^T\mathbf{J}^{-1}(2\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ &+ \mathbf{z}_{p'}^T\mathbf{J}^{-1}(2\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\mathbf{J}^{-1}\frac{\partial \mathbf{z}_p}{\partial z_{pr}}\Bigg)\end{aligned}$$

for $p \neq p'$. Since $k(\mathbf{z}_p, \mathbf{z}_{p'}) = 1$, this second derivative will be 0 for $p = p'$. For the same reason, the derivatives of $\mathbb{E}_{p(\mathbf{b}_n|\boldsymbol{\mu},\boldsymbol{\Sigma}_n)}[k(\mathbf{b}_n, \mathbf{b}_n)]$ with respect to z_{pr}, and $\mathbf{c}_r$ and $\mathbf{J}_{ij}$ are also all 0. The gradients with respect to $\boldsymbol{\mu}$ are

$$\frac{\partial \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}_n)}[k(\mathbf{z}_p, \boldsymbol{\mu})]}{\partial \boldsymbol{\mu}} = \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}_n)}[k(\mathbf{z}_p, \boldsymbol{\mu})]\Big(-(\mathbf{J} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\mu} + \mathbf{z}^T\mathbf{J}^{-1}(\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\Big).$$

and

$$\begin{aligned}\frac{\partial \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)}[k(\mathbf{z}_p, \boldsymbol{\mu})k(\mathbf{z}_{p'}, \boldsymbol{\mu})]}{\partial \boldsymbol{\mu}} =& \mathbb{E}_{p(\boldsymbol{\mu}|\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)}[k(\mathbf{z}_p, \boldsymbol{\mu})k(\mathbf{z}_{p'}, \boldsymbol{\mu})] \\ &\times \Big(-(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(2\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu} \\ &+ (\mathbf{z}_p + \mathbf{z}_{p'})^T\mathbf{J}^{-1}(2\mathbf{J}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\Big).\end{aligned}$$

Given the equations above, the derivatives of $\mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{ZB}}]$ and $\mathbb{E}_{Q_{AB}}[\mathbf{K}_{\mathbf{ZB}}\mathbf{K}_{\mathbf{ZB}}^T]$ with respect to $\mathbf{c}_r$ could simply be taken using $b_{nr} = \mathbf{k}_{\mathbf{X}_{ir}\mathbf{x}_n}^T \mathbf{K}_{\mathbf{X}_{ir}\mathbf{X}_{ir}}^{-1} \mathbf{c}_r$ and

$$\frac{\partial \mathbf{b}_n}{\partial c_{pr}} = \mathbf{e}_p^T \mathbf{K}_{\mathbf{X}_{ir}\mathbf{X}_{ir}}^{-1} \mathbf{c}_r$$

together with the chain rule. Here, $\mathbf{e}_p$ is a $P \times 1$ vector whose pth entry is 1 and other entries are 0. The integer P stands for the number of inducing points.

Finally, the gradients with respect to the hyperparameter $\mathbf{J}_{ij}$ are

$$\begin{aligned}\frac{\partial \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[k(\mathbf{z}_p,\boldsymbol{\mu})]}{\partial \mathbf{J}_{ij}} =& \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[k(\mathbf{z}_p,\boldsymbol{\mu})]\Bigg(-\frac{1}{2}|\mathbf{J}^{-1}\boldsymbol{\Sigma}+\mathbf{I}|^{-\frac{3}{2}} tr\Big((\mathbf{J}^{-1}\boldsymbol{\Sigma}+\mathbf{I})\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\Big) \\ & -\frac{1}{2}\mathbf{z}^T\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\mathbf{z} - \mathbf{z}^T\mathbf{J}^{-1}(\mathbf{J}^{-1}+\boldsymbol{\Sigma}^{-1})\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\mathbf{z} \\ & -\frac{1}{2}\mathbf{z}^T\mathbf{J}^{-1}\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\mathbf{J}^{-1}\mathbf{z} - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\boldsymbol{\Sigma}\mathbf{J}^{-1}\boldsymbol{\mu} \\ & + \mathbf{z}^T\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}(\mathbf{J}^{-1}\boldsymbol{\Sigma}+\mathbf{I})\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{z}^T\mathbf{J}^{-1}\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Bigg),\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[k(\mathbf{z}_p,\boldsymbol{\mu})k(\mathbf{z}_{p'},\boldsymbol{\mu})]}{\partial \mathbf{J}_{ij}} &= \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})}[k(\mathbf{z}_p,\boldsymbol{\mu})k(\mathbf{z}_{p'},\boldsymbol{\mu})]\times \\ & \Bigg(-\frac{1}{2}|2\mathbf{J}^{-1}\boldsymbol{\Sigma}+\mathbf{I}|^{-\frac{3}{2}} tr\Big((2\mathbf{J}^{-1}\boldsymbol{\Sigma}+\mathbf{I})\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\Big) - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\boldsymbol{\Sigma}\mathbf{J}^{-1}\boldsymbol{\mu} \\ & -\frac{1}{2}\mathbf{z}_p^T\Big(\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}} + 2\mathbf{J}^{-1}(\mathbf{J}^{-1}+\boldsymbol{\Sigma}^{-1})\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}} + \mathbf{J}^{-1}\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\mathbf{J}^{-1}\Big)\mathbf{z}_p \\ & -\frac{1}{2}\mathbf{z}_{p'}^T\Big(\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}} + 2\mathbf{J}^{-1}(\mathbf{J}^{-1}+\boldsymbol{\Sigma}^{-1})\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}} + \mathbf{J}^{-1}\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\mathbf{J}^{-1}\Big)\mathbf{z}_{p'} \\ & + (\mathbf{z}_p+\mathbf{z}_{p'})^T\Big(\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}(\mathbf{J}^{-1}\boldsymbol{\Sigma}+\mathbf{I})\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{J}^{-1}\frac{\partial \mathbf{J}^{-1}}{\partial \mathbf{J}_{ij}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\Big)\Bigg).\end{aligned}$$