Supplement to "Nonparametric Prior Learning in Differential Equation Modeling"

Contents

A Additional Results

A.1 PAC-Bayesian for Inverse Problems

The present work focuses on the infinite-dimensional Bayesian statistical approach for solving inverse problems associated with partial differential equations (PDEs). Bayesian statistics have a deep connection with the PAC-Bayesian framework, as illustrated in Germain et al. (2016), specifically in the finite-dimensional setting. Although deep connections have been observed in the field of machine learning, to our knowledge, there seems to be little discussion on solving inverse problems of PDEs by linking the Bayesian inverse approach with PAC-Bayesian learning. For the reader's convenience, we intend to provide a brief introduction to the PAC-Bayesian learning theory from the perspective of infinite-dimensional inverse problems.

As in the main text, let us assume that the forward operator $\mathcal{G} : \mathcal{U} \times \mathcal{X} \to \mathcal{H}$ with $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{H}$ being separable Banach spaces. We assume that the data are generated by:

$$y_j = \mathcal{G}(u, x_j) + \eta_j, \quad \forall\, j = 1, \ldots, m, \tag{1}$$

where u represents the model parameter, $x_j \in \mathcal{X}$ is the measurement point or an input function, $y_j \in \mathcal{Y}$ is the measured data, and η_j is the independent identically distributed (i.i.d.) random noise. In the following, we assume that the forward operator can be written as $\mathcal{L}_{x_j}(\mathcal{G}(u))$, where $\mathcal{G}(u)$ belongs to some separable Hilbert space and $\mathcal{L}_{x_j}$ is a bounded linear operator determined by $x_j \in \mathcal{X}$. Denote the vector $\boldsymbol{x} = (x_1, \cdots, x_m)^T$, the vector $\boldsymbol{y} = (y_1, \cdots, y_m)^T$, and the vector $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_m)^T$, then model (1) can be rewritten in a more compact form

$$\boldsymbol{y} = \mathcal{L}_{\boldsymbol{x}}\mathcal{G}(u) + \boldsymbol{\eta},$$

where $\mathcal{L}_{\boldsymbol{x}}\mathcal{G}(u) := (\mathcal{L}_{x_1}\mathcal{G}(u), \ldots, \mathcal{L}_{x_m}\mathcal{G}(u))^T$. For solving inverse problems, we assume that there is a background true parameter denoted by $u^\dagger$. Then, the forward model will be

$$\boldsymbol{y} = \mathcal{L}_{\boldsymbol{x}}\mathcal{G}(u^\dagger) + \boldsymbol{\eta}.$$

Define $\mathcal{P}(\mathcal{Y})$ as the set of all probability measures defined on a separable Banach space $\mathcal{Y}$. In the field of infinite-dimensional Bayesian inverse theory, we assume that the unknown parameter u is a random function distributed according to a prior probability measure $\mathbb{P}$, which is an element of $\mathcal{P}(\mathcal{U})$. Then, we introduce a posterior probability measure $\mathbb{Q}$, also an element of $\mathcal{P}(\mathcal{U})$, such that

$$\frac{d\mathbb{Q}}{d\mathbb{P}}(u) = \frac{1}{Z_m} \exp\left(-\sum_{j=1}^{m} \Phi(u; y_j) \right), \tag{2}$$

where $\Phi(u; y_j)$ is the potential function derived from the likelihood function, and Z_m is the normalization constant, defined as

$$Z_m := \int_{\mathcal{U}} \exp\left(-\sum_{j=1}^{m} \Phi(u; y_j) \right) \mathbb{P}(du).$$

For a fixed parameter u, the observed data vector $\boldsymbol{y}$ is determined by the input vector $\boldsymbol{x}$ and the noise vector $\boldsymbol{\eta}$. As in the main text, we also express the potential function $\Phi(u; y_j)$ as $\Phi(u; x_j, \eta_j)$ for $j = 1, \ldots, m$, and assume that the formula (2) is well defined.

Assume that there is a probability measure $\mathbb{D}(dx, dy) := \mathbb{D}_1(dx)\mathbb{D}_2(dy)$, with $\mathbb{D}_1$ and $\mathbb{D}_2$ being two probability measures defined on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Define $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, then we have $z := (x, y) \sim \mathbb{D} \in \mathcal{P}(\mathcal{Z})$. For convenience, let us denote the dataset as $S = \{z_j = (x_j, y_j)\}_{j=1}^m$, consisting of i.i.d. samples from the distribution $\mathbb{D}$, that is, $S \sim \mathbb{D}^m$. Obviously, the measure $\mathbb{D}_2$ depends on the background true parameter $u^\dagger$, the distribution $\mathbb{D}_1$, and the distribution of the noise, denoted by $\mathbb{D}_3$ (a probability measure defined on $\mathcal{Y}$). Here, we assume that $\{x_j\}_{j=1}^m$ are generated according to the probability measure $\mathbb{D}_1$. To provide a brief illustration of the PAC-Bayesian framework, we introduce a loss function $\ell : \mathcal{U} \times \mathcal{Z} \to \mathbb{R}$. Accordingly, we aim to minimize the expected error under the data distribution $\mathbb{D}$, that is,

$$\mathcal{L}(u, \mathbb{D}) := \mathbb{E}_{z\sim\mathbb{D}}\ell(u, z).$$

For solving inverse problems, the true distributions of u and the noise $\boldsymbol{\eta}$ are unknown, i.e., the data distribution $\mathbb{D}$ is unknown. Hence, it is necessary to define the empirical error

$$\hat{\mathcal{L}}(u, S) := \frac{1}{m} \sum_{j=1}^{m} \ell(u, z_j).$$

To give the key bound in the PAC-Bayesian framework, we define the so-called Gibbs error as $\mathcal{L}(\mathbb{Q}, \mathbb{D}) := \mathbb{E}_{u\sim\mathbb{Q}}\mathcal{L}(u, \mathbb{D})$ and its empirical counterpart as $\hat{\mathcal{L}}(\mathbb{Q}, S) := \mathbb{E}_{u\sim\mathbb{Q}}\hat{\mathcal{L}}(u, S)$. Let us denote $D_{\mathrm{KL}}(\cdot||\cdot)$ as the usual Kullback-Leibler (KL) divergence between two probability measures. Now, we can state the following theorem, which provides an upper bound for the unknown generalization error based on its empirical estimate.

Theorem A.1. *Given data distribution $\mathbb{D}$, parameter space $\mathcal{U}$, loss function $\ell(u, z)$, prior measure $\mathbb{P} \in \mathcal{P}(\mathcal{U})$, confidence level $\delta \in (0, 1]$, and $\beta > 0$, with probability at least $1 - \delta$ over samples $S \sim \mathbb{D}^m$, we have that, for all $\mathbb{Q} \in \mathcal{P}(\mathcal{U})$:*

$$\mathcal{L}(\mathbb{Q}, \mathbb{D}) \leq \hat{\mathcal{L}}(\mathbb{Q}, S) + \frac{1}{\beta}\left[D_{KL}(\mathbb{Q}||\mathbb{P}) + \ln\frac{1}{\delta} + \Psi(\beta, m)\right], \tag{3}$$

where $\Psi(\beta, m) := \ln \mathbb{E}_{u\sim\mathbb{P}}\mathbb{E}_{S\sim\mathbb{D}^m} \exp\left[\beta\left(\mathcal{L}(u, \mathbb{D}) - \hat{\mathcal{L}}(u, S)\right)\right]$ is a term that depends on the tail of the distribution of the loss function.

If $\beta = m$, given the loss function $\ell(u, z) := \Phi(u, x, \eta)$, and with a proper estimate of $\Psi(\beta, m)$, then the minimum value on the right-hand side of estimate (3) corresponds to the posterior measure specified in equation (2). Generally, the optimal solution provided by the PAC-Bayesian bound may not correspond to the posterior measure given by the Bayes' formula, and the optimal solution typically corresponds to the posterior measure obtained from a generalized Bayes' formula. The generalization bound (3) is selected from Germain et al. (2016) and Alquier et al. (2016), which generalizes the original PAC-Bayes bounds (McAllester 1998). For a comprehensive illustration, we refer to Alquier (2024).

Remark A.2. *For the current work, we focus on the problem of learning data-dependent prior measures. Theorem A.1 can be conceptually adapted to the statistical inverse problems of PDEs. Bounding the term $\Psi(\beta, m)$ and generalizing the results shown in Alquier et al. (2016) for inverse problems present interesting future research challenges. Similar to the adaptation of the nonparametric Bayesian approach to inverse problems (Nickl 2020, Nickl & Söhl 2017, Monard et al. 2021), such a generalization is not a trivial task, due to issues of measurability, appropriate estimations of PDEs, and the singularity of infinite-dimensional measures.*

A.2 Data-Independent Prior

In the main text, we provide illustrations of the data-dependent prior case. For completeness, we intend to give a detailed explanation of the data-independent prior case, i.e., the prior measure $\mathbb{P}_S^\theta$ is assumed to be independent of the dataset S. Hence, we denote this prior measure as $\mathbb{P}^\theta$.

Case 1: Bounded loss. For simplicity, let us first assume that the loss function $\ell(u, z)$ is bounded by (a, b), where $-\infty < a < b < +\infty$. Under this assumption, we can apply Hoeffding's lemma to obtain the following corollary.

Corollary A.3. *Assume that all of the assumptions in Theorem 1 hold true and the loss function $\ell(u, \boldsymbol{z})$ is bounded in $[a, b]$ with $-\infty < a < b < +\infty$. For any confidence level $\delta \in (0, 1]$, $\gamma \geq 2\sqrt{n}$, and $\lambda \geq 2\sqrt{n}$, the inequality*

$$\begin{aligned}\mathcal{L}(\mathscr{Q}, \mathcal{T}) \leq \hat{\mathcal{L}}(\mathscr{Q}, S_1, \ldots, S_n) &+ \left(\frac{1}{\lambda} + \frac{1}{\gamma}\right) D_{KL}(\mathscr{Q}||\mathscr{P}) \\ &+ \frac{1}{\gamma}\sum_{i=1}^{n} \mathbb{E}_{\theta\sim\mathscr{Q}} D_{KL}\big(\mathbb{Q}(S_i, \mathbb{P}^\theta)||\mathbb{P}^\theta\big) + \left(\frac{\gamma}{8nm} + \frac{\lambda}{8n}\right)(b - a)^2 + \frac{1}{\sqrt{n}}\ln\frac{1}{\delta}\end{aligned}$$

holds uniformly over all hyperposteriors $\mathscr{Q} \in \mathcal{P}(\Theta)$ with probability $1 - \delta$.

Proof. Remember that we assumed the loss function $\ell(u, z)$ is bounded in $[a, b]$ with $-\infty < a < b < +\infty$. Under this assumption, we can apply Hoeffding's lemma (Lemma D.1 in Mohri et al. (2018)) to each factor of the term $\mathbb{E}\big[\exp(\sqrt{n}\Pi^1(\gamma) + \sqrt{n}\Pi^2(\lambda))\big]$, obtaining:

$$\begin{aligned}\mathbb{E}\left[e^{(\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda))}\right] \leq & \left\{\prod_{i=1}^{n}\mathbb{E}_{u_i\sim\mathbb{P}^\theta}\left[e^{\frac{\gamma^2}{8n^2m}(b-a)^2}\right]\right\}^{\frac{\sqrt{n}}{\gamma}}\left\{e^{\frac{\lambda^2}{8n}(b-a)^2}\right\}^{\frac{\sqrt{n}}{\lambda}} \\ & \leq \exp\left(\left(\frac{\gamma}{8\sqrt{n}m}+\frac{\lambda}{8\sqrt{n}}\right)(b-a)^2\right),\end{aligned} \tag{4}$$

when $\gamma \geq 2\sqrt{n}$ and $\lambda \geq 2\sqrt{n}$. Plugging this estimate into the general estimate shown in Theorem 1 of the main text, we obtain an explicit bound on the moment generating function, which completes the proof. □

Case 2: Unbound loss. The inverse problems of PDEs can be seen as PDE-constrained regression problems, which are often modeled using unbounded loss functions, such as the squared loss function. To provide appropriate estimates of the log moment generating term, we consider sub-Gaussian type loss functions, which are employed in the study of regression problems with a fixed finite number of parameters (Germain et al. 2016).

Assumption A.4 (sub-Gaussian assumption related to Π^1)**.** *For $i = 1, \ldots, n$, we assume that the random variables $\mathcal{L}(u_i, \mathbb{D}_i) - \ell(u_i, z_i)$ are sub-Gaussian with variance factor s_I^2, i.e., for all $\tilde{\gamma} > 0$, we have*

$$V_i^1 := \mathbb{E}_{\mathbb{D}_i\sim\mathcal{T}}\mathbb{E}_{z_i\sim\mathbb{D}_i}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}^\theta}\exp\left(\tilde{\gamma}\left[\mathcal{L}(u_i, \mathbb{D}_i) - \ell(u_i, z_i)\right]\right) \leq \exp\left(\frac{\tilde{\gamma}^2 s_I^2}{2}\right).$$

Assumption A.5 (sub-Gaussian assumption related to Π^2)**.** *For $i = 1, \ldots, n$, we assume that the random variables $\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathcal{L}(\mathbb{Q}(S, \mathbb{P}^\theta), \mathbb{D}) - \mathcal{L}(\mathbb{Q}(S_i, \mathbb{P}^\theta), \mathbb{D}_i)$ are sub-Gaussian with variance factor s_{II}^2, i.e., for all $\tilde{\lambda} > 0$, we have*

$$\begin{aligned}V_i^2 := & \mathbb{E}_{\mathbb{D}_i\sim\mathcal{T}}\mathbb{E}_{S_i\sim\mathbb{D}_i^m}\mathbb{E}_{\theta\sim\mathscr{P}}\exp\left(\tilde{\lambda}\left[\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathcal{L}(\mathbb{Q}(S, \mathbb{P}^\theta), \mathbb{D}) - \mathcal{L}(\mathbb{Q}(S_i, \mathbb{P}^\theta), \mathbb{D}_i)\right]\right) \\ & \leq \exp\left(\frac{\tilde{\lambda}^2 s_{II}^2}{2}\right).\end{aligned}$$

With these two assumptions, we can prove the following corollary.

Corollary A.6. *Assume that all of the assumptions in Theorem 1 and Assumptions A.4 and A.5 hold true. In addition, we assume $2\sqrt{n} \leq \lambda$ and $2\sqrt{n} \leq \gamma$. For any confidence level $\delta \in (0, 1]$, the inequality*

$$\begin{aligned}\mathcal{L}(\mathscr{Q}, \mathcal{T}) \leq & \hat{\mathcal{L}}(\mathscr{Q}, S_1, \ldots, S_n) + \left(\frac{1}{\lambda}+\frac{1}{\gamma}\right)D_{KL}(\mathscr{Q}||\mathscr{P}) \\ & + \frac{1}{\gamma}\sum_{i=1}^{n}\mathbb{E}_{\theta\sim\mathscr{Q}}D_{KL}\big(\mathbb{Q}(S_i, \mathbb{P}^\theta)||\mathbb{P}^\theta\big) + \frac{\gamma s_I^2}{2nm} + \frac{\lambda s_{II}^2}{2n} + \frac{1}{\sqrt{n}}\ln\frac{1}{\delta}\end{aligned}$$

holds uniformly over all hyperposteriors $\mathscr{Q} \in \mathcal{P}(\Theta)$ with probability $1 - \delta$.

Proof. In the following $\frac{\gamma}{nm}$ will be recognized as $\tilde{\gamma}$ and we assume $\gamma \geq 2\sqrt{n}$ and $\lambda \geq 2\sqrt{n}$. Employing the assumption of V_i^1 in Assumption A.4 of the main text, we have

$$\begin{aligned}
&\mathbb{E}\left\{\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_1\sim\mathbb{P}^\theta}\cdots\mathbb{E}_{u_n\sim\mathbb{P}^\theta}\exp\left(\frac{\gamma}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}[\mathcal{L}(u_i,\mathbb{D}_i)-\ell(u_i,z_{ij})]\right)\right\}^{\frac{2\sqrt{n}}{\gamma}} \\
&\leq\left\{\mathbb{E}\prod_{i=1}^{n}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}^\theta}\prod_{j=1}^{m}\exp\left(\frac{\gamma}{nm}[\mathcal{L}(u_i,\mathbb{D}_i)-\ell(u_i,z_{ij})]\right)\right\}^{\frac{2\sqrt{n}}{\gamma}} \\
&\leq\exp\left(\frac{\gamma s_{\mathrm{I}}^2}{\sqrt{n}m}\right),
\end{aligned} \tag{5}$$

which can be written more compactly as follow:

$$\left(\mathbb{E}\exp(2\sqrt{n}\Pi^1(\gamma))\right)^{1/2}\leq\exp\left(\frac{\gamma s_{\mathrm{I}}^2}{2\sqrt{n}m}\right). \tag{6}$$

Employing the assumption of V_i^2 in Assumption A.5 of the main text, we have

$$\begin{aligned}
&\mathbb{E}\left\{\mathbb{E}_{\theta\sim\mathscr{P}}\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathcal{L}(\mathbb{Q}(S,\mathbb{P}^\theta),\mathbb{D})-\mathcal{L}(\mathbb{Q}(S_i,\mathbb{P}^\theta),\mathbb{D}_i)\right]\right)\right\}^{\frac{2\sqrt{n}}{\lambda}} \\
&\leq\left\{\prod_{i=1}^{n}\mathbb{E}\mathbb{E}_{\theta\sim\mathscr{P}}\exp\left(\frac{\lambda}{n}\left[\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathcal{L}(\mathbb{Q}(S,\mathbb{P}^\theta),\mathbb{D})-\mathcal{L}(\mathbb{Q}(S_i,\mathbb{P}^\theta),\mathbb{D}_i)\right]\right)\right\}^{2\frac{\sqrt{n}}{\lambda}} \\
&\leq\left\{\prod_{i=1}^{n}\exp\left(\frac{\lambda^2 s_{\mathrm{II}}^2}{2n^2}\right)\right\}^{\frac{2\sqrt{n}}{\lambda}}=\exp\left(\frac{\lambda s_{\mathrm{II}}^2}{\sqrt{n}}\right),
\end{aligned}$$

which yields

$$\left(\mathbb{E}\exp(2\sqrt{n}\Pi^2(\lambda))\right)^{1/2}\leq\exp\left(\frac{\lambda s_{\mathrm{II}}^2}{2\sqrt{n}}\right). \tag{7}$$

Considering estimates (**??**) and (**??**), we obtain the desired result. □

Remark A.7. *In Assumptions A.4 and A.5, we introduced the parameters $\tilde{\gamma}$ and $\tilde{\lambda}$, which correspond to the parameters γ/nm and λ/n in Corollary A.6. Typically, we set $\gamma = n\beta$ with $\beta = \sqrt{m}$ or $\beta = m$ when $2\sqrt{n} \leq n\beta$, and $\lambda = 2\sqrt{n}$ or n when $2\sqrt{n} \leq n$. For deriving Corollary A.6, we require $\lambda \geq 2\sqrt{n}$ and $\gamma \geq 2\sqrt{n}$, conditions not mentioned in the work of meta-learning (Rothfuss et al. 2021) for the finite-dimensional case. Since the random variables $\Pi^1(\gamma)$ and $\Pi^2(\lambda)$ are not independent, we cannot obtain*

$$\mathbb{E}\left[e^{(\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda))}\right]=\mathbb{E}\left[e^{\sqrt{n}\Pi^1(\gamma)}\right]\mathbb{E}\left[e^{\sqrt{n}\Pi^2(\lambda)}\right].$$

Hence, it may be necessary for us to employ

$$\begin{aligned}
\mathbb{E}\left[e^{(\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda))}\right]&\leq\left(\mathbb{E}\left[e^{2\sqrt{n}\Pi^1(\gamma)}\right]\right)^{1/2}\left(\mathbb{E}\left[e^{2\sqrt{n}\Pi^2(\lambda)}\right]\right)^{1/2} \\
&\leq\left(\mathbb{E}\left[e^{\gamma\Pi^1(\gamma)}\right]\right)^{\sqrt{n}/\gamma}\left(\mathbb{E}\left[e^{\lambda\Pi^2(\lambda)}\right]\right)^{\sqrt{n}/\lambda},
\end{aligned}$$

which requires that $\lambda \geq 2\sqrt{n}$ and $\gamma \geq 2\sqrt{n}$.

A.3 Results Under Sub-Gamma Assumptions

In the main text, we provide the sub-Gaussian assumptions for the terms Π^1 and Π^2 related to the loss function. In the following, we list the sub-Gamma assumptions, which often appear in the studies of the PAC-Bayesian theory (Germain et al. 2016, Rothfuss et al. 2021).

Assumption A.8 (sub-Gamma assumption related to Π^1)**.** *For $i = 1, \ldots, n$, we assume that the random variables*

$$\mathcal{L}(u_i, \mathbb{D}_i) - \ell(u_i, z_i)$$

are sub-Gamma with variance factor s_I^2 and scale parameter $c_I < 1$, i.e.,

$$\begin{aligned} V_i^1 :=& \mathbb{E}_{\mathbb{D}_i \sim \mathcal{T}} \mathbb{E}_{z_i \sim \mathbb{D}_i} \mathbb{E}_{\theta \sim \mathscr{P}} \mathbb{E}_{u_i \sim \mathbb{P}^\theta} \exp\left(\tilde{\gamma}\left[\mathcal{L}(u_i, \mathbb{D}_i) - \ell(u_i, z_i)\right]\right) \\ \leq& \exp\left(\frac{\tilde{\gamma}^2 s_I^2}{2(1 - \tilde{\gamma} c_I)}\right) \end{aligned}$$

for all $\tilde{\gamma} \in (0, 1/c_I)$.

Assumption A.9 (sub-Gamma assumption related to Π^2)**.** *For $i = 1, \ldots, n$, we assume that the random variables*

$$\mathbb{E}_{\mathbb{D} \sim \mathcal{T}} \mathbb{E}_{S \sim \mathbb{D}^m} \mathcal{L}(\mathbb{Q}(S, \mathbb{P}^\theta), \mathbb{D}) - \mathcal{L}(\mathbb{Q}(S_i, \mathbb{P}^\theta), \mathbb{D}_i)$$

are sub-Gamma with variance factor s_{II}^2 and scale parameter $c_{II} < 1$, i.e.,

$$\begin{aligned} V_i^2 :=& \mathbb{E}_{\mathbb{D}_i \sim \mathcal{T}} \mathbb{E}_{S_i \sim \mathbb{D}_i^m} \mathbb{E}_{\theta \sim \mathscr{P}} \exp\left(\tilde{\lambda}\left[\mathbb{E}_{\mathbb{D} \sim \mathcal{T}} \mathbb{E}_{S \sim \mathbb{D}^m} \mathcal{L}(\mathbb{Q}(S, \mathbb{P}^\theta), \mathbb{D}) - \mathcal{L}(\mathbb{Q}(S_i, \mathbb{P}^\theta), \mathbb{D}_i)\right]\right) \\ \leq& \exp\left(\frac{\tilde{\lambda}^2 s_{II}^2}{2(1 - \tilde{\lambda} c_{II})}\right) \end{aligned}$$

for all $\tilde{\lambda} \in (0, 1/c_{II})$.

With these two assumptions, we can derive the following two corollaries of Theorem 1 in the main text, when we assume that the prior measures are either data-independent or data-dependent, respectively.

Corollary A.10. *Assume that all of the assumptions in Theorem 1 and Assumptions A.8 and A.9 hold true. In addition, we assume $2\sqrt{n} \leq \lambda \leq n$ and $2\sqrt{n} \leq \gamma \leq nm$. For any confidence level $\delta \in (0, 1]$, the inequality*

$$\begin{aligned} \mathcal{L}(\mathscr{Q}, \mathcal{T}) \leq& \hat{\mathcal{L}}(\mathscr{Q}, S_1, \ldots, S_n) + \left(\frac{1}{\lambda} + \frac{1}{\gamma}\right) D_{KL}(\mathscr{Q} || \mathscr{P}) \\ &+ \frac{1}{\gamma} \sum_{i=1}^n \mathbb{E}_{\theta \sim \mathscr{Q}} D_{KL}\big(\mathbb{Q}(S_i, \mathbb{P}^\theta) || \mathbb{P}^\theta\big) \\ &+ \frac{\gamma s_I^2}{2nm\left(1 - c_I \frac{\gamma}{nm}\right)} + \frac{\lambda s_{II}^2}{2n\left(1 - \frac{\lambda}{n} c_{II}\right)} + \frac{1}{\sqrt{n}} \ln \frac{1}{\delta} \end{aligned}$$

holds uniformly over all hyper-posteriors $\mathscr{Q} \in \mathcal{P}(\Theta)$ with probability $1 - \delta$.

Proof. In the following $\frac{\gamma}{nm}$ will be recognized as $\tilde{\gamma}$ and we assume $\gamma \geq 2\sqrt{n}$ and $\lambda \geq 2\sqrt{n}$. Employing the assumption of V_i^1 in Assumption A.8, we have

$$\begin{aligned}
&\mathbb{E}\left\{\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_1\sim\mathbb{P}^\theta}\cdots\mathbb{E}_{u_n\sim\mathbb{P}^\theta}\exp\left(\frac{\gamma}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}[\mathcal{L}(u_i,\mathbb{D}_i)-\ell(u_i,z_{ij})]\right)\right\}^{\frac{2\sqrt{n}}{\gamma}} \\
&\leq\left\{\mathbb{E}\prod_{i=1}^{n}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}^\theta}\prod_{j=1}^{m}\exp\left(\frac{\gamma}{nm}[\mathcal{L}(u_i,\mathbb{D}_i)-\ell(u_i,z_{ij})]\right)\right\}^{\frac{2\sqrt{n}}{\gamma}} \\
&\leq\exp\left(\frac{\gamma s_{\mathrm{I}}^2}{\sqrt{n}m\left(1-\frac{\gamma}{nm}c_{\mathrm{I}}\right)}\right),
\end{aligned} \tag{8}$$

which can be written more compactly as follow:

$$\left(\mathbb{E}\exp(2\sqrt{n}\Pi^1(\gamma))\right)^{1/2}\leq\exp\left(\frac{\gamma s_{\mathrm{I}}^2}{2\sqrt{n}m\left(1-\frac{\gamma}{nm}c_{\mathrm{I}}\right)}\right). \tag{9}$$

Employing the assumption of V_i^2 in Assumption A.9, we have

$$\begin{aligned}
&\mathbb{E}\left\{\mathbb{E}_{\theta\sim\mathscr{P}}\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathcal{L}(\mathbb{Q}(S,\mathbb{P}^\theta),\mathbb{D})-\mathcal{L}(\mathbb{Q}(S_i,\mathbb{P}^\theta),\mathbb{D}_i)\right]\right)\right\}^{\frac{2\sqrt{n}}{\lambda}} \\
&\leq\left\{\prod_{i=1}^{n}\mathbb{E}\mathbb{E}_{\theta\sim\mathscr{P}}\exp\left(\frac{\lambda}{n}\left[\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathcal{L}(\mathbb{Q}(S,\mathbb{P}^\theta),\mathbb{D})-\mathcal{L}(\mathbb{Q}(S_i,\mathbb{P}^\theta),\mathbb{D}_i)\right]\right)\right\}^{2\frac{\sqrt{n}}{\lambda}} \\
&\leq\left\{\prod_{i=1}^{n}\exp\left(\frac{\lambda^2 s_{\mathrm{II}}^2}{2n^2\left(1-\frac{\lambda}{n}c_{\mathrm{II}}\right)}\right)\right\}^{2\frac{\sqrt{n}}{\lambda}}=\exp\left(\frac{\lambda s_{\mathrm{II}}^2}{\sqrt{n}\left(1-\frac{\lambda}{n}c_{\mathrm{II}}\right)}\right),
\end{aligned}$$

which yields

$$\left(\mathbb{E}\exp(2\sqrt{n}\Pi^2(\lambda))\right)^{1/2}\leq\exp\left(\frac{\lambda s_{\mathrm{II}}^2}{2\sqrt{n}\left(1-\frac{\lambda}{n}c_{\mathrm{II}}\right)}\right). \tag{10}$$

Considering estimates (**??**) and (**??**), we obtain the desired result. □

Corollary A.11. *Assume that all of the assumptions in Theorem 1, Assumptions A.8, A.9, and 1 (in the main text) hold true. In addition, we assume* $2\sqrt{n}\leq\lambda\leq n$ *and* $4\sqrt{n}\leq 2\gamma\leq nm$. *For any confidence level* $\delta\in(0,1]$, *the inequality*

$$\begin{aligned}
\mathcal{L}(\mathscr{Q},\mathcal{T})\leq\hat{\mathcal{L}}(\mathscr{Q},S_1,\ldots,S_n)&+\left(\frac{1}{\lambda}+\frac{1}{\gamma}\right)D_{KL}(\mathscr{Q}||\mathscr{P}) \\
&+\frac{1}{\gamma}\sum_{i=1}^{n}\mathbb{E}_{\theta\sim\mathscr{Q}}D_{KL}\big(\mathbb{Q}(S_i,\mathbb{P}_{S_i}^\theta)||\mathbb{P}_{S_i}^\theta\big)+\frac{n}{2\gamma}\ln\Psi_E \\
&+\frac{\gamma s_I^2}{nm\left(1-c_I\frac{\gamma}{nm}\right)}+\frac{\lambda s_{II}^2}{2n\left(1-\frac{\lambda}{n}c_{II}\right)}+\frac{1}{\sqrt{n}}\ln\frac{1}{\delta}
\end{aligned}$$

holds uniformly over all hyper-posteriors $\mathscr{Q}\in\mathcal{P}(\Theta)$ *with probability* $1-\delta$.

Proof. For the term V_i^1, we have

$$\begin{aligned} V_i^1 =& \mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}_{S_i}^{\theta}}\exp\left(\tilde{\gamma}\left[\mathcal{L}(u_i,\mathbb{D}_i)-\ell(u_i,z_i)\right]\right) \\ \leq& \mathbb{E}_{S_i^{'}\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}_{S_i^{'}}^{\theta}}\Psi(S_i,S_i^{'},u_i,\theta)\exp\left(\tilde{\gamma}[\mathcal{L}(u_i,\mathbb{D}_i)-\ell(u_i,z_i)]\right) \\ \leq& \Psi_E^{1/2}\left(\mathbb{E}_{S_i^{'}\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}_{S_i^{'}}^{\theta}}\exp\left(2\tilde{\gamma}[\mathcal{L}(u_i,\mathbb{D}_i)-\ell(u_i,z_i)]\right)\right)^{1/2} \\ \leq& \Psi_E^{1/2}\exp\left(\frac{\tilde{\gamma}^2 s_{\mathrm{I}}^2}{1-\tilde{\gamma}c_{\mathrm{I}}}\right), \end{aligned} \tag{11}$$

where we need $0 < 2\tilde{\gamma} < 1/c_{\mathrm{I}}$ for deriving the last inequality. Through some simple calculations as for the bounded loss case, we find that

$$\left(\mathbb{E}e^{2\sqrt{n}\Pi^1(\gamma)}\right)^{1/2} \leq \Psi_E^{\frac{n\sqrt{n}}{2\gamma}}\exp\left(\frac{\gamma s_{\mathrm{I}}^2}{\sqrt{n}m\left(1-\frac{\gamma}{nm}c_{\mathrm{I}}\right)}\right). \tag{12}$$

Combining estimates on the term $\mathbb{E}e^{\sqrt{n}\Pi^2(\lambda)}$ (not given here since it is similar to the data-independent case), we obtain

$$\mathbb{E}\left[e^{\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda)}\right] \leq \Psi_E^{\frac{n\sqrt{n}}{2\gamma}}\exp\left(\frac{\gamma s_{\mathrm{I}}^2}{\sqrt{n}m\left(1-\frac{\gamma}{nm}c_{\mathrm{I}}\right)}+\frac{\lambda s_{\mathrm{II}}^2}{2\sqrt{n}\left(1-\frac{\lambda}{n}c_{\mathrm{II}}\right)}\right). \tag{13}$$

Combining estimate (**??**) with the results given in Theorem 1, we obtain the desired result. □

A.4 Discussion of Assumption 1

In this subsection, we revisit the example provided in Subsection 2.3, which demonstrates that Assumption 1 from the main text could hold true under specific conditions. In the main text, we posit that the space $\mathcal{U}$ is a separable Hilbert space, and the prior measure is defined as $\mathbb{P}_S^{\theta} := \mathcal{N}(f(S;\theta),\mathcal{C}_0)$, where $\mathcal{C}_0$ is a positive definite, symmetric trace class operator serving as the covariance of the prior measure. Let $\{e_k\}_{k=1}^{\infty}$ be an orthogonal basis for $\mathcal{U}$. Then, we assume that

$$\mathcal{C}_0 = \sum_{k=1}^{\infty}\lambda_k^2 e_k\otimes e_k,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots$. For the mean function $f(S;\theta)$, we assume that it belongs to the space $\mathcal{U}^2$ and all of $f(S;\theta)$ are restricted in a ball $B_{\mathcal{U}^2}(R_u)$ defined as follow:

$$B_{\mathcal{U}^2}(R_u) := \{f\in\mathcal{U}^2 :\ \|\mathcal{C}_0^{-1}f\|_{\mathcal{U}} \leq R_u\}, \tag{14}$$

where R_u is a prespecified large enough positive constant. For the above condition (5), similar requirements are employed in classical investigations of inverse problems of partial differential equations (PDEs). When studying inverse problems of PDEs using regularization methods, it is important for us to derive conditional stability estimates, which require

the introduction of an admissible set (Engl et al. 1996). The admissible set is usually defined as a ball in some appropriate Banach space. For example, in the investigation of the backward diffusion problem, the admissible set is defined as a ball in the Sobolev space $H^{2\epsilon}$ (for any fixed $\epsilon > 0$) with a fixed finite radius (Li et al. 2009). In studies of the inverse coefficient for steady-state Darcy flow equations, the admissible set is defined as a ball in the C^1 space with a finite radius (Richter 1981, Vollmer 2013). In some recent investigations on statistical inverse problems, similar admissible sets have also been introduced; see Section 2 of Nickl (2023), for example. From a practical point of view, it is reasonable to restrict the parameter of interest within a ball of some Banach space, since such a priori information can often be obtained in practical problems, e.g., we approximately know the range of the wave propagation velocity for the underground medium in seismic imaging.

Under these assumptions, we have

$$\begin{aligned}\Psi(S, S', u, \theta) = \exp\Bigg(& \langle \mathcal{C}_0^{-1}(f(S;\theta) - f(S';\theta)), u - f(S;\theta)\rangle_{\mathcal{U}} \\ & - \frac{1}{2}\|\mathcal{C}_0^{-1/2}(f(S;\theta) - f(S';\theta))\|_{\mathcal{U}}^2 \Bigg)\end{aligned} \tag{15}$$

based on Theorem 2.23 in Prato & Zabczyk (2014). For a small real number $0 < \epsilon < \frac{1}{4\lambda_1^2}$, we have the following estimate

$$\Psi(S, S', u, \theta)^2 \leq T_1 \cdot T_2,$$

where

$$T_1 = \exp\left(\frac{1}{\epsilon}\|\mathcal{C}_0^{-1}(f(S;\theta) - f(S';\theta))\|_{\mathcal{U}}^2\right), \quad T_2 = \exp\left(\epsilon\|u - f(S;\theta)\|_{\mathcal{U}}^2\right).$$

Obviously, we know that

$$\begin{aligned}\Psi_E \leq & \mathbb{E}_{S'\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{S\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{u\sim\mathbb{P}^{\theta}_{S'}} T_1 \cdot T_2 \\ \leq & \mathbb{E}_{S'\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{S\sim\mathbb{D}_{\mathcal{T}}} T_1 \cdot \exp\left(2\epsilon\|f(S';\theta) - f(S;\theta)\|_{\mathcal{U}}^2\right)\mathbb{E}_{u\sim\mathbb{P}_0}\exp\left(2\epsilon\|u\|_{\mathcal{U}}^2\right) \\ \leq & e^{8\epsilon R_u^2}\left[\det\left(\mathrm{I} - 4\epsilon\mathcal{C}_0\right)\right]^{-\frac{1}{2}}\mathbb{E}_{S'\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{S\sim\mathbb{D}_{\mathcal{T}}} T_1 \\ = & e^{8\epsilon R_u^2}\prod_{k=1}^{\infty}(1 - 4\epsilon\lambda_k^2)^{-\frac{1}{2}}\mathbb{E}_{S'\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{S\sim\mathbb{D}_{\mathcal{T}}} T_1,\end{aligned}$$

and where $\mathbb{P}_0 := \mathcal{N}(0, \mathcal{C}_0)$. Remember the bounded condition of $f(S;\theta)$ for any S and θ, we finally obtain

$$\Psi_E \leq e^{\frac{(4+8\epsilon^2)R_u^2}{\epsilon}}\prod_{k=1}^{\infty}(1 - 4\epsilon\lambda_k^2)^{-\frac{1}{2}} < +\infty, \tag{16}$$

with $0 < \epsilon < \frac{1}{4\lambda_1^2}$.

Remark A.12. *Since we have not assumed that the parameter u is uniformly bounded, the term $\Psi(S, S', u, \theta)$ is not generally bounded. Therefore, we cannot verify the α-differentially private condition under the current setting. The right-hand side of inequality (7) contains a term $e^{\frac{(4+8\epsilon^2)R_u^2}{\epsilon}}$, which can become very large even for moderate values of R_u. However, it is only the term $\ln\Psi_E$ that appears in the generalization bound and scales like R_u^2.*

A.5 Discussion of Bounded Loss

Before further discussions, we note that the following corollary, which concerns the generalization bound under a bounded loss assumption and data-dependent prior, holds true.

Corollary A.13. *Assume that all of the assumptions in Theorem 1 and Assumption 1 hold true. In addition, we assume that the loss function $\ell(u, \boldsymbol{z})$ is bounded in $[a, b]$ with $-\infty < a < b < +\infty$. For any confidence level $\delta \in (0, 1]$, $\gamma \geq 2\sqrt{n}$, and $\lambda \geq 2\sqrt{n}$, the inequality*

$$\mathcal{L}(\mathscr{Q}, \mathcal{T}) \leq \hat{\mathcal{L}}(\mathscr{Q}, S_1, \ldots, S_n) + \left(\frac{1}{\lambda} + \frac{1}{\gamma}\right) D_{KL}(\mathscr{Q}||\mathscr{P}) + \frac{1}{\gamma}\sum_{i=1}^{n} \mathbb{E}_{\theta\sim\mathscr{Q}} D_{KL}\big(\mathbb{Q}(S_i, \mathbb{P}_{S_i}^{\theta})||\mathbb{P}_{S_i}^{\theta}\big) + \frac{n}{2\gamma}\ln\Psi_E + \left(\frac{\gamma}{4nm} + \frac{\lambda}{8n}\right)(b-a)^2 + \frac{1}{\sqrt{n}}\ln\frac{1}{\delta}$$

holds uniformly over all hyperposteriors $\mathscr{Q} \in \mathcal{P}(\Theta)$ with probability $1 - \delta$.

Proof. Since $\mathbb{E}\mathbb{E}_{\theta\sim\mathscr{P}}\cdot = \mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}\cdot$ (The expectation $\mathbb{E}$ is defined as in Theorem 1 of the main text), the estimates of $e^{2\sqrt{n}\Pi^2(\lambda)}$ will be the same as for the data-independent case. In the following, we concentrate on the estimates of $e^{2\sqrt{n}\Pi^1(\gamma)}$. For $\gamma \geq 2\sqrt{n}$, simple calculations yield

$$\left(\mathbb{E}e^{2\sqrt{n}\Pi^1(\gamma)}\right)^{1/2} \leq (\mathbb{E}T_1)^{\frac{\sqrt{n}}{\gamma}}, \tag{17}$$

where

$$T_1 = \mathbb{E}_{\theta\sim\mathscr{P}} \prod_{i=1}^{n} \mathbb{E}_{u_i\sim\mathbb{P}_{S_i}^{\theta}} \exp\left(\frac{\gamma}{nm}\sum_{j=1}^{m}[\mathbb{E}_{z\sim\mathbb{D}_i}\ell(u_i, z) - \ell(u_i, z_{ij})]\right).$$

For the term $\mathbb{E}T_1$ in inequality (8), we have

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\mathscr{P}} \prod_{i=1}^{n} \mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}} \mathbb{E}_{u_i\sim\mathbb{P}_{S_i}^{\theta}} \exp\left(\frac{\gamma}{nm}\sum_{j=1}^{m}[\mathbb{E}_{z\sim\mathbb{D}_i}\ell(u_i, z) - \ell(u_i, z_{ij})]\right) \\
=&\mathbb{E}_{\theta\sim\mathscr{P}} \prod_{i=1}^{n} \mathbb{E}_{S_i'\sim\mathbb{D}_{\mathcal{T}}} \mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}} \mathbb{E}_{u_i\sim\mathbb{P}_{S_i}^{\theta}} \exp\left(\frac{\gamma}{nm}\sum_{j=1}^{m}[\mathbb{E}_{z\sim\mathbb{D}_i}\ell(u_i, z) - \ell(u_i, z_{ij})]\right) \\
\leq&\mathbb{E}_{\theta\sim\mathscr{P}} \prod_{i=1}^{n} \mathbb{E}_{S_i'\sim\mathbb{D}_{\mathcal{T}}} \mathbb{E}_{u_i\sim\mathbb{P}_{S_i'}^{\theta}} \mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}} \Psi(S_i, S_i', u_i, \theta) \exp\left(\frac{\gamma}{nm}\sum_{j=1}^{m}[\mathbb{E}_{z\sim\mathbb{D}_i}\ell(u_i, z) - \ell(u_i, z_{ij})]\right) \\
\leq&\mathbb{E}_{\theta\sim\mathscr{P}} \prod_{i=1}^{n} \left\{\mathbb{E}_{S_i'\sim\mathbb{D}_{\mathcal{T}}} \mathbb{E}_{u_i\sim\mathbb{P}_{S_i'}^{\theta}} \mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}} \Psi(S_i, S_i', u_i, \theta)^2\right\}^{1/2} \\
&\qquad \left\{\mathbb{E}_{u_i\sim\mathbb{P}_{S_i'}^{\theta}} \mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}} \exp\left(\frac{2\gamma}{nm}\sum_{j=1}^{m}[\mathbb{E}_{z\sim\mathbb{D}_i}\ell(u_i, z) - \ell(u_i, z_{ij})]\right)\right\}^{1/2},
\end{aligned}$$

which implies

$$\mathbb{E}T_1 \leq \Psi_E^{\frac{n}{2}} e^{\frac{\gamma^2}{4nm}(b-a)^2} \tag{18}$$

with Ψ_E defined as in Assumption 1 of the main text. Plugging estimate (**??**) into estimate (8), we find that

$$\left(\mathbb{E}e^{2\sqrt{n}\Pi^1(\gamma)}\right)^{1/2} \leq \Psi_E^{\frac{n\sqrt{n}}{2\gamma}} e^{\frac{\gamma}{4\sqrt{n}m}(b-a)^2}. \tag{19}$$

Combining estimates on the term $\mathbb{E}e^{2\sqrt{n}\Pi^2(\lambda)}$ (not given here since it is similar to the data-independent case), we obtain

$$\mathbb{E}\left[e^{(\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda))}\right] \leq \Psi_E^{\frac{n\sqrt{n}}{2\gamma}} \exp\left(\left(\frac{\gamma}{4\sqrt{n}m} + \frac{\lambda}{8\sqrt{n}}\right)(b-a)^2\right). \tag{20}$$

Combining estimate (**??**) with the results given in Theorem 1 of the main text, we complete the proof. □

Inspired by investigations on the generalized Bayes' method, where the likelihood function is not directly derived from some probability distributions (Germain et al. 2016, Guedj 2019), we introduce the following truncated loss function:

$$\ell(u,z) := \begin{cases} \dfrac{1}{2}\|\mathcal{L}_x\mathcal{G}(u) - y\|_{\mathcal{H}}^2, & \text{if } \dfrac{1}{2}\|\mathcal{L}_x\mathcal{G}(u) - y\|_{\mathcal{H}}^2 \leq N, \\ N, & \text{if } \dfrac{1}{2}\|\mathcal{L}_x\mathcal{G}(u) - y\|_{\mathcal{H}}^2 > N, \end{cases} \tag{21}$$

Here, N is a prespecified positive constant. With this assumption, the loss function is contained within the interval $[0, N]$, which allows us to apply Theorems A.3 and A.13 by replacing $(b-a)^2$ with N^2. Introducing such a truncation may seem unnatural at first glance, especially from the perspective of Bayesian inverse methods. However, in the following, we provide an explanation that illustrates the appropriateness of introducing such a truncated loss function.

As discussed in Appendix A.4, an admissible set—usually defined as a ball in some Banach space—will be introduced when we discuss the conditional stability estimates of inverse problems (Li et al. 2009, Richter 1981, Vollmer 2013). Based on this consideration, we assume that there exists a constant R_u such that

$$\operatorname{supp} \mathscr{E} \subset B_{\mathcal{U}}(R_u)$$

with $\mathscr{E} \in \mathcal{P}(\mathcal{U})$ and $B_{\mathcal{U}}(R_u) := \{u \in \mathcal{U} : \|u\|_{\mathcal{U}} \leq R_u\}$. Assume the forward operator $\mathcal{G}$ is continuous that satisfies

$$\begin{aligned} \|\mathcal{L}_x\mathcal{G}(u)\|_{\mathcal{H}} \leq & M(\|u\|_{\mathcal{U}}), \\ \|\mathcal{L}_x\mathcal{G}(u) - \mathcal{L}_x\mathcal{G}(v)\|_{\mathcal{H}} \leq & \tilde{M}(\|u\|_{\mathcal{U}}, \|v\|_{\mathcal{U}})\|u - v\|_{\mathcal{U}}, \end{aligned}$$

where $x \in \mathcal{X}$, $M(\cdot)$ is a monotonic non-decreasing function, and $\tilde{M}(\cdot,\cdot)$ is a function monotonic non-decreasing separately in each argument. By our assumptions on $\mathcal{E}$, the parameter $u \in B_{\mathcal{U}}(R_u)$ almost surely, which indicates

$$\begin{aligned} \|\mathcal{L}_x\mathcal{G}(u)\|_{\mathcal{H}} \leq & M(R_u), \\ \|\mathcal{L}_x\mathcal{G}(u) - \mathcal{L}_x\mathcal{G}(v)\|_{\mathcal{H}} \leq & \tilde{M}(R_u, R_u)\|u - v\|_{\mathcal{U}} \end{aligned}$$

for all $u, v \in B_{\mathcal{U}}(R_u)$.

In the study of regularization methods for inverse problems, the noise is assumed to be bounded by a constant, known as the "noise level" (Engl et al. 1996). Intuitively, the noise level cannot be too large, or it would obscure the true measured data. Based on this consideration, we assume that the noise $\eta \in \mathcal{H}$ is bounded by $\mathcal{L}_x\mathcal{G}(u)$, i.e.,

$$\|\eta\|_{\mathcal{H}} \leq C_\eta \|\mathcal{L}_x\mathcal{G}(u)\|_{\mathcal{H}}$$

with C_η be a positive constant. Usually, the constant C_η is assumed to be smaller than 1. Let us define the loss function as follow:

$$\ell(u, z) = \frac{1}{2}\|\mathcal{L}_x\mathcal{G}(u) - y\|_{\mathcal{H}}^2.$$

Then, we easily know that

$$\begin{aligned}\ell(u, z) \leq & \|\mathcal{L}_x\mathcal{G}(u) - \mathcal{L}_x\mathcal{G}(u^\dagger)\|_{\mathcal{H}}^2 + \|\eta\|_{\mathcal{H}}^2 \\ \leq & \tilde{M}(R_u, R_u)^2\|u - u^\dagger\|_{\mathcal{U}}^2 + C_\eta^2 M(R_u)^2\|u^\dagger\|_{\mathcal{U}}^2 \\ \leq & \left(4\tilde{M}(R_u, R_u)^2 + C_\eta^2 M(R_u)^2\right) R_u^2.\end{aligned}$$

Hence, we can take the constant N in formula (9) as follow:

$$N := \left(4\tilde{M}(R_u, R_u)^2 + C_\eta^2 M(R_u)^2\right) R_u^2.$$

Under the current setting, the loss function may not have a probabilistic interpretation. Hence we can only formulate a generalized Bayes' formula that is well accepted in the studies of PAC-Bayesian learning theory.

A.6 Analysis of the Backward Diffusion and Darcy Flow Problems

In this section, let us provide detailed illustrations of the two concrete inverse problems of PDEs that were mentioned in Subsection 3.2 of the main text.

The backward diffusion problem. We apply the theory to the backward diffusion problem, which is one of the most investigated inverse problems of PDEs (Jia et al. 2016, Jia, Peng, Gao & Li 2018, Stuart 2010). Let $\Omega \subset \mathbb{R}^d (d \leq 3)$ be a bounded open set with smooth boundary $\partial\Omega$. We define the Hilbert space $\mathcal{H}$ and the operator A as follows:

$$\mathcal{H} = (L^2(\Omega), \langle\cdot,\cdot\rangle, \|\cdot\|), \quad A = -\Delta, \quad \mathcal{D}(A) = H^2(\Omega) \cap H_0^1(\Omega).$$

Then the forward diffusion equation is an ordinary differential equation in $\mathcal{H}$:

$$\frac{dv}{dt} + Av = 0, \quad v(0) = u.$$

Elements of the backward diffusion problem:

- Datasets: the solution of the diffusion equation at time $T > 0$, i.e., $v(T)$ is the measured data;

- Interested parameter: the initial function u of the diffusion equation.

By the operator semigroup theory, we know that the solution at time T is $v(T) = \exp(-AT)u$, i.e., the forward operator $\mathcal{G}(u) = \exp(-AT)u$ and $\mathcal{L}_x := \mathrm{Id}$ is the identity operator from $\mathcal{H}$ to $\mathcal{H}$. Let $\mathcal{U} = \mathcal{H}$, i.e., the interested parameter u is assumed to belong to $\mathcal{H}$. For $j = 1, \ldots, m$, we have

$$y_j = \mathcal{G}(u, x_j) + \eta_j = \exp(-AT)u + \eta_j, \tag{22}$$

where $\{x_j\}_{j=1}^m$ are dummy variables. The above problem (10) can be written compactly as

$$\boldsymbol{y} = \mathcal{L}_{\boldsymbol{x}}\mathcal{G}(u) + \boldsymbol{\eta} = \exp(-AT)\boldsymbol{u} + \boldsymbol{\eta}, \tag{23}$$

where $\boldsymbol{u} = (u_1, \ldots, u_m)^T$ with $u_j = u$ for $j = 1, \ldots, m$.

For this example, we take $\Gamma := \tau^2\mathrm{Id}$ with $\tau \in \mathbb{R}^+$, $\mathcal{C}_1 = A^{-2}$, $\mathcal{C}_0 = \lambda A^{-2}$ ($\lambda > 0$) and $\alpha = 0$. Concerned with $\mathcal{C}_0$, we have the following eigen-system decomposition:

$$\mathcal{C}_0 e_k = \lambda_k e_k, \quad \forall\, k = 1, 2, \ldots,$$

where $\{\lambda_k\}_{k=1}^\infty$ are arranged in a descending order and $\{e_k\}_{k=1}^\infty$ are normalized eigenfunctions. According to the properties of the operator A, we know that $\lambda_k \asymp k^{-\frac{4}{d}}$, which yields $s_0 = s_1 = \frac{d}{4}$. In fact, for every $s > s_0$, we have

$$\mathrm{Tr}(\mathcal{C}_0^s) = \sum_{k=1}^\infty \lambda_k^s \asymp \sum_{k=1}^\infty k^{-\frac{4s}{d}} < +\infty.$$

With these discussions, we can take $\frac{d}{4} < s = \tilde{s} < 1$ for the backward diffusion problem discussed here. In addition, the parameter $\alpha = 0$ and $\mathcal{Y} = \mathcal{H}^{-s}$.

Now we give estimates required in Assumptions 2 of the main text (replace the assumptions on the forward operator by inequalities (32) of the main text) for the current problem as follows:

$$\begin{aligned}
\|\mathcal{C}_1^{\frac{s}{2}}\Gamma^{-\frac{1}{2}}u\|_{\mathcal{H}} =& \tau^{-1}\|\mathcal{C}_1^{\frac{s}{2}}u\|_{\mathcal{H}},\\
\|\mathcal{C}_1^{-\frac{\rho}{2}}\Gamma^{\frac{1}{2}}u\|_{\mathcal{H}} =& \tau\|\mathcal{C}_1^{-\frac{\rho}{2}}u\|_{\mathcal{H}},\\
\|\mathcal{C}_1^{-\frac{s}{2}}\Gamma^{-\frac{1}{2}}e^{-AT}u\|_{\mathcal{H}} \le& \tau^{-1}\sup_k\left[\lambda^{\frac{s}{2}}\lambda_k^{\frac{1}{2}-s}e^{-T\sqrt{\frac{\lambda}{\lambda_k}}}\right]\|u\|_{\mathcal{U}^{1-\tilde{s}}},\\
\|\mathcal{C}_1^{-\frac{s}{2}}\Gamma^{-\frac{1}{2}}e^{-AT}(u_1-u_2)\|_{\mathcal{H}} \le& \tau^{-1}\sup_k\left[\lambda^{\frac{s}{2}}\lambda_k^{\frac{1}{2}-s}e^{-T\sqrt{\frac{\lambda}{\lambda_k}}}\right]\|u_1-u_2\|_{\mathcal{U}^{1-\tilde{s}}},
\end{aligned} \tag{24}$$

which indicate that $C_1 = \tau^{-1}$, $C_2 = \tau$, and $M_1 = M_2$ in inequalities (32) of the main text equal to $\tau^{-1}\sup_k\left[\lambda^{\frac{s}{2}}\lambda_k^{\frac{1}{2}-s}e^{-T\sqrt{\frac{\lambda}{\lambda_k}}}\right]$. Inequalities (**??**) are easily verified by employing the standard estimates of parabolic equations (see for example the Section 1.2 of Dashti & Stuart (2017) or the book Pazy (1983)). For reader's convenience, we provide the proof of (**??**) in Appendix B.2.

In order to employ Lemma 1 of the main text, we need to verify the condition $2\tilde{\gamma}\tilde{C}^2M_2^2 \le \lambda_1^{-1}$. After a simple calculation similar to the proof of (**??**), we find that

$$2\lambda_1^{1+\frac{s}{2}}\sup_k\left[\lambda_k^{\frac{1}{2}-\frac{s}{2}}\left(\frac{\lambda}{\lambda_k}\right)^{s/2}e^{-T\sqrt{\frac{\lambda}{\lambda_k}}}\right] \le \tau\lambda^{s/2} \tag{25}$$

ensures the above condition when choosing $\tilde{\gamma} = 1$. By assuming (**??**) and employing estimates (**??**), we find out the parameters s_{I}^2 and s_{II}^2 in Lemma 1 of the main text are as follows:

$$\begin{aligned} s_{\mathrm{I}}^2 =& \frac{\lambda_1^s}{\lambda^s\tau^2}\sup_k\left[\lambda^s\lambda_k^{1-2s}e^{-2T\sqrt{\frac{\lambda}{\lambda_k}}}\right]\left(8R_u^2+\mathrm{Tr}(\mathcal{C}_1^s)+\frac{\lambda_1^s\sup_k\left[\lambda^s\lambda_k^{1-2s}e^{-2T\sqrt{\frac{\lambda}{\lambda_k}}}\right]}{4\lambda^s\tau^2}\right) \\ & -\frac{1}{2}\ln\det\left(\mathrm{Id}-2\frac{\lambda_1^s}{\lambda^2\tau^2}\sup_k\left[\lambda^s\lambda_k^{1-2s}e^{-2T\sqrt{\frac{\lambda}{\lambda_k}}}\right]\mathcal{C}_0\right), \\ s_{\mathrm{II}}^2 =& \frac{\lambda_1^s}{\tau^2}\sup_k\left[\lambda_k^{2-4s}e^{-2\sqrt{\frac{\lambda}{\lambda_k}}T}\right]\left(12R_u^2+\mathrm{Tr}(\mathcal{C}_0)+\frac{3\beta}{m}\mathrm{Tr}(\mathcal{C}_0^s)\right)+12R_u^2 \\ & +\frac{\lambda_1^{2s}}{4\tau^4}\sup_k\left[\lambda_k^{4-8s}e^{-4\sqrt{\frac{\lambda}{\lambda_k}}T}\right]R_u^4. \end{aligned} \tag{26}$$

To maintain a clear illustration, we provide the calculation details of (**??**) in Appendix B.2. Using the above formula, we can immediately derive the PAC-Bayesian bounds given in Corollaries A.6 and 1 of the main text for the data-independent prior and data-dependent prior cases when $\gamma = nm$ and $\lambda = n$.

We choose $\gamma = nm$ and $\lambda = n$ for the purpose of simplifying the analysis. However, with the current techniques, we are unable to provide estimates when $\gamma = n\sqrt{m}$ and $\lambda = 2\sqrt{n}$, especially when the base prior measure is a Gaussian measure (not truncated). If the base prior measure is assumed to be a truncated Gaussian measure, we can obtain explicit estimations. The derivation is similar to the illustration given in the following for nonlinear problems. Developing an appropriate PAC-Bayesian theory for the case of $\gamma = n\sqrt{m}$ and $\lambda = 2\sqrt{n}$ with a Gaussian base prior measure (not truncated) remains an open problem for future work.

The Darcy flow problem. Regarding the linear forward operator case, we apply the general theory to a steady-state Darcy flow problem, which is a well-known example in the field of statistical inverse problems (Dashti & Stuart 2017, Jia et al. 2022). Let $\Omega \subset \mathbb{R}^2$ be a bounded open set with a smooth boundary $\partial\Omega$, and define $\mathcal{U}$ as the Hilbert space $(L^2(\Omega), \langle\cdot,\cdot\rangle, \|\cdot\|)$. The Darcy flow equation has the following form:

$$\begin{aligned} -\nabla\cdot(e^u\nabla w) &= f \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{27}$$

where f denotes the sources, and $e^{u(x)}$ describes the permeability of the porous medium.

Elements of the inverse problems for Darcy flow:

- Datasets: At some discrete points $\{x_j\}_{j=1}^m$, we measure the value of the solution w through a measurement operator defined as follow:

$$\mathcal{L}_{x_j}(w) = \int_\Omega \frac{1}{2\pi\delta^2}e^{-\frac{1}{2\delta^2}\|x-x_j\|^2}w(x)dx,$$

  with $\delta > 0$ being a sufficiently small number and $x_j \in \Omega$ for $j = 1, \ldots, m$.

- Interested parameter: The function $u(x)$.

For this problem, the operator $\mathcal{G}$ is the solution operator of the Darcy flow. Hence, we have

$$y_j = \mathcal{L}_{x_j}\mathcal{G}(u) + \eta_j = \mathcal{L}_{x_j}(w) + \eta_j,$$

which can be written compactly as

$$\boldsymbol{y} = \mathcal{L}_{\boldsymbol{x}}\mathcal{G}(u) + \boldsymbol{\eta}.$$

Obviously, the space $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$ in this example. Let us set $\Gamma := \tau^2(\tau \in \mathbb{R}^+)$ and the covariance operator of the prior measure $\mathcal{C}_0 := A^{-3}$, where $A := \alpha(\mathrm{Id} - \Delta)(\alpha > 0)$ with the domain of Δ given by $\mathcal{D}(\Delta) := \left\{u \in H^2(\Omega) : \frac{\partial u}{\partial \boldsymbol{n}} = 0\right\}$. Under this setting, we know that $s_0 = \frac{1}{3}$, i.e., we can take $\tilde{s} > \frac{1}{3}$. Since $\mathcal{H}$ is a finite-dimensional space, we need not introduce the operator $\mathcal{C}_1$. Now we give estimates required in Assumptions 2 of the main text for the Darcy flow problem as follows:

$$\begin{aligned} |\mathcal{L}_x\mathcal{G}(u)| \leq & M_3(\|u\|_{\mathcal{U}^{1-\tilde{s}}}), \\ |\mathcal{L}_x\mathcal{G}(u_1) - \mathcal{L}_x\mathcal{G}(u_2)| \leq & M_4(\|u_1\|_{\mathcal{U}^{1-\tilde{s}}}, \|u_2\|_{\mathcal{U}^{1-\tilde{s}}})\|u_1 - u_2\|_{\mathcal{U}^{1-\tilde{s}}}, \end{aligned} \tag{28}$$

where

$$\begin{aligned} M_3(\|u\|_{\mathcal{H}^{1-\tilde{s}}}) = & \frac{2d_\Omega^2}{\pi\delta}\exp\left(\|u\|_{\mathcal{U}^{1-\tilde{s}}}\right)\|f\|_{\mathcal{U}}, \\ M_4(\|u_1\|_{\mathcal{H}^{1-\tilde{s}}}, \|u_2\|_{\mathcal{U}^{1-\tilde{s}}}) = & \frac{2d_\Omega^2}{\pi\delta}\exp\left(\|u_1\|_{\mathcal{U}^{1-\tilde{s}}} + \|u_2\|_{\mathcal{U}^{1-\tilde{s}}}\right)\|f\|_{\mathcal{U}}. \end{aligned}$$

Here d_Ω denotes the diameter of the domain Ω. It is not difficult to prove the above results since these estimates are essentially the same as those of Theorem 17 proved in Jia et al. (2022) (a more careful calculation of the constants is needed). Concerned with $\lim_{x\to x'}|\mathcal{L}_x\mathcal{G}(u) - \mathcal{L}_{x'}\mathcal{G}(u)| = 0$, it can be easily verified by employing similar techniques used for proving estimates (12). So we omit these proofs for simplicity.

With estimates (12) at hand, we find out the parameters s_{I}^2 and s_{II}^2 in Lemma 2 are as follows:

$$s_{\mathrm{I}}^2 = \frac{100d_\Omega^8}{\pi^4\delta^4}e^{8R_u} + \frac{4d_\Omega^4}{\pi^2\delta^2}e^{2R_u}, \quad s_{\mathrm{II}}^2 = \frac{400d_\Omega^8}{\pi^4\delta^4}e^{4R_u}. \tag{29}$$

Relying on the above formulas (**??**), we immediately obtain the PAC-Bayesian bounds presented in Theorems A.6 and 1 of the main text for both the data-independent prior and data-dependent prior cases. Unlike the backward diffusion problem, here we do not restrict the parameters γ and λ to specific values, such as $\gamma = nm$ and $\lambda = n$, since there are no $\tilde{\gamma}$ and $\tilde{\lambda}$ appearing in s_{I}^2 and s_{II}^2 of Lemma 2 in the main text. As a preliminary study, our main aim is to develop a general theoretical framework that can yield explicit bounds. For deriving sharper estimates, more detailed structural information of the PDEs is expected to be utilized. This suggests that different techniques for estimation may need to be developed for inverse problems associated with various PDEs.

Here, we provide two typical examples that satisfy our general formulation. It is worth mentioning that our general formulation can also be applied to study other problems, such as the inverse problems of fractional differential equations (Jia et al. 2017, Jia, Peng, Gao & Li 2018, Jin & Rundell 2015), or the inverse medium scattering problems (Bao et al. 2015, Jia et al. 2019, Jia, Yue, Peng & Gao 2018).

A.7 Details of Constructing Learning Algorithm

Here, let us provide more detailed discussions on constructing learning algorithms based on the PAC-Bayesian generalization bound derived in the main text. Restricted to finite-dimensional space, the following illustrations obviously hold true as illustrated in Rothfuss et al. (2021) for machine learning problems. For the inverse problems of PDEs, we seek base generalized posterior measures that are highly interpretable. Hence, we select the parameters $\gamma = nm$, $\lambda = n$, and base posterior measures based on the classical Bayes' formula, introduced formally in (2) of the main text or rigorously in (6) with $\beta = m$. In the following, we denote $Z_m = Z_m(S, \mathbb{P}_S^\theta)$, since Z_m depends on the dataset and the prior measure. With this choice of the base posterior measure, the Kullback-Leibler (KL) divergence term of the base prior and posterior measures in PAC-Bayesian bounds (e.g., estimates in Corollaries A.13 and 1 of the main text) can be further reduced as follows:

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{Q}(S_i, \mathbb{P}_{S_i}^\theta)||\mathbb{P}_{S_i}^\theta) &= \int_{\mathcal{U}} \left(\ln \frac{d\mathbb{Q}(S_i, \mathbb{P}_{S_i}^\theta)}{d\mathbb{P}_{S_i}^\theta} \right) d\mathbb{Q}(S_i, \mathbb{P}_{S_i}^\theta) \\ &= -\ln Z_m(S_i, \mathbb{P}_{S_i}^\theta) - \int_{\mathcal{U}} \sum_{j=1}^m \Phi(u; z_{ij}) d\mathbb{Q}(S_i, \mathbb{P}_{S_i}^\theta) \\ &= -\ln Z_m(S_i, \mathbb{P}_{S_i}^\theta) - \mathbb{E}_{u\sim\mathbb{Q}(S_i, \mathbb{P}_{S_i}^\theta)} \left[\sum_{j=1}^m \Phi(u; z_{ij}) \right]. \end{aligned} \tag{30}$$

Choosing the loss function $\ell(u; z)$ to be the potential function $\Phi(u; z)$ and using (**??**), we immediately find

$$\begin{aligned} \frac{1}{nm} \sum_{i=1}^n \mathbb{E}_{\theta\sim\mathscr{Q}} D_{\mathrm{KL}}(\mathbb{Q}(S_i, \mathbb{P}_{S_i}^\theta)||\mathbb{P}_{S_i}^\theta) = &-\frac{1}{nm} \sum_{i=1}^n \mathbb{E}_{\theta\sim\mathscr{Q}} \ln Z_m(S_i, \mathbb{P}_{S_i}^\theta) \\ &- \hat{\mathcal{L}}(\mathscr{Q}, S_1, \ldots, S_n). \end{aligned} \tag{31}$$

Plugging the equality (**??**) into the PAC-Bayesian bounds, e.g., the estimate given in Corollary 1 of the main text, we obtain that

$$\begin{aligned} \mathcal{L}(\mathscr{Q}, \mathcal{T}) \leq &-\frac{1}{nm} \sum_{i=1}^n \mathbb{E}_{\theta\sim\mathscr{Q}} \ln Z_m(S_i, \mathbb{P}_{S_i}^\theta) + \frac{m+1}{nm} D_{\mathrm{KL}}(\mathscr{Q}||\mathscr{P}) \\ &+ \frac{1}{2m} \ln \Psi_E + \frac{s_{\mathrm{I}}^2}{2} + \frac{s_{\mathrm{II}}^2}{2} + \frac{1}{\sqrt{n}} \ln \frac{1}{\delta} \end{aligned} \tag{32}$$

holds uniformly with probability $1 - \delta$. To minimize the right-hand side of (**??**), we need to solve the following optimization problem

$$\arg\min_{\mathscr{Q}\in\mathcal{P}(\Theta)} \mathbb{E}_{\theta\sim\mathscr{Q}} \left[-\frac{1}{m+1} \sum_{i=1}^n \ln Z_m(S_i, \mathbb{P}_{S_i}^\theta) \right] + D_{\mathrm{KL}}(\mathscr{Q}||\mathscr{P}). \tag{33}$$

In addressing this optimization problem, we present Theorem 3 in the main text, which provides an analytic formula for the optimal measure $\mathscr{Q}$. In the following remark, we offer a discussion on the utility of simplifying the problem and deriving the Bayes' formula as presented in Theorem 3 of the main text.

Remark A.14. *Inspired by the work of Rothfuss et al. (2021) for finite-dimensional machine learning model, we choose the base posterior measure to be the posterior measure given by Bayes' formula. Relying on this, we obtain the hyper-posterior measure analytically. Generally, the bounds we derived in the present work (e.g., estimates in Corollaries A.13 and 1 of the main text) yield an optimization problem as follow:*

$$\arg\min_{\mathscr{Q}\in\mathcal{P}(\Theta)} \hat{\mathcal{L}}(\mathscr{Q}, S_1, \ldots, S_n) + \frac{m+1}{nm} D_{KL}(\mathscr{Q}||\mathscr{P}) + \frac{1}{nm}\sum_{i=1}^{n} \mathbb{E}_{\theta\sim\mathscr{Q}} D_{KL}(\mathbb{Q}(S_i, \mathbb{P}^{\theta}_{S_i})||\mathbb{P}^{\theta}_{S_i}),$$

where the base posterior measure $\mathbb{Q}(S_i, \mathbb{P}^{\theta}_{S_i})$ is not necessarily provided by the Bayes' formula. Solving an optimization problem like the above one turns into a difficult two-level optimization problem (Amit & Meir 2018, Pentina & Lampert 2014) which can hardly be used for the inverse problems of PDEs due to the large computational complexity for solving PDEs.

With the formula (13) in the main text, we are ready to extract information about the parameter θ from the hyper-posterior measure $\mathscr{Q}$. There are three types of methods listed as follows:

- Evaluate the maximum a posteriori estimator using gradient-based optimization algorithms (Bottou et al. 2018, Ito & Jin 2015);
- Sample from the hyper-posterior measure using Markov chain Monte Carlo type algorithms, e.g., preconditioned Crank-Nicolson algorithm (Cotter et al. 2013, Dashti & Stuart 2017);
- Approximate sampling algorithms, such as variational inference methods, e.g., the mean-field based variational inference method (Jia et al. 2021), and the Stein variational gradient descent method (Jin & Zou 2010, Liu & Wang 2016).

Notice that the log-likelihood function of the hyperposerior measure given in Theorem 3 of the main text is defined as follows:

$$\sum_{i=1}^{n} \ln Z_m(S_i, \mathbb{P}^{\theta}_{S_i}) = \sum_{i=1}^{n} \ln \int_{\mathcal{U}} \exp\Big(-\sum_{j=1}^{m} \Phi(u; z_{ij})\Big) \mathbb{P}^{\theta}_{S_i}(du). \tag{34}$$

For the backward diffusion problem considered in Appendix A.6, it is possible to find the explicit formula of (13). However, for general inverse problems of PDEs, such as the Darcy flow problem illustrated in Appendix A.6, it is generally not possible to calculate explicitly. Generally, formula (13) can be calculated by

$$\sum_{i=1}^{n} \ln Z_m(S_i, \mathbb{P}^{\theta}_{S_i}) \approx \sum_{i=1}^{n} \ln \left[\frac{1}{L} \sum_{\ell=1}^{L} \exp\Big(-\sum_{j=1}^{m} \Phi(u^i_\ell; z_{ij})\Big)\right], \tag{35}$$

where $\{u^i_\ell\}_{\ell=1}^{L}$ are samples from the base prior measure $\mathbb{P}^{\theta}_{S_i}$ for $i = 1, \ldots, n$. By applying a simple reduction, we obtain the following result:

$$\sum_{i=1}^{n} \ln Z_m(S_i, \mathbb{P}^{\theta}_{S_i}) \approx \sum_{i=1}^{n} \ln \left[\sum_{\ell=1}^{L} \exp\Big(-\sum_{j=1}^{m} \Phi(u^i_\ell; z_{ij})\Big)\right] - n \ln L. \tag{36}$$

To evaluate the right-hand side of (**??**), we need to calculate nL forward PDEs, which is a time-consuming procedure (it is computationally more feasible compared with the two-level optimization problem illustrated in Remark A.14). Hence, we focus on evaluating the maximum a posteriori estimate in the present work. For particular linear problems, e.g., backward diffusion, we can compute Z_m much more efficiently, which allows us to utilize approximate sampling or Markov chain Monte Carlo type methods.

A.8 Numerical Details and More Results

We establish our theories within the framework of separable infinite-dimensional function spaces, which will be useful for constructing discretization-invariant algorithms. With this goal in mind, it is desirable that the employed neural network exhibits a discretization-invariant property, meaning that the network is capable of learning a nonlinear operator rather than merely a nonlinear function. Currently, there are several works focused on constructing neural networks for nonlinear operator learning (Anandkumar et al. 2020, Bhattacharya et al. 2021, Li et al. 2020, Nelsen & Stuart 2021). In our numerical studies, we employ the Fourier neural operator (FNO), as proposed in Li et al. (2020), which consists of three Fourier layers. We choose the FNO due to its simplicity in implementation and its computational efficiency. For readers interested in more detailed theoretical studies on FNO, we refer to Kovachki et al. (2021).

In our setting, the measurement data are discrete variables that are equally spaced in a statistical sense, as illustrated in Subsection 2.1. However, these data cannot be used directly as inputs for the FNO, since the FNO operates on functions as both inputs and outputs. To address this issue, we apply the adjoint measurement operator to the noisy data to obtain a function, and then we use the resulting function evaluated at the grid points as the input for the FNO. Now, let us provide some details.

As in the main text, let $\{\phi_k\}_{k=1}^N$ be the basis functions of the finite element discretization. Here, we use boldface letters to denote vectors and matrices. Following Bui-Thanh et al. (2013), we define the mass matrix $\boldsymbol{M} = (M_{k\ell})_{k,\ell=1,\dots,N}$ by

$$M_{k\ell} = \int_\Omega \phi_k(\boldsymbol{x})\phi_\ell(\boldsymbol{x})d\boldsymbol{x}. \tag{37}$$

We define the weighted finite-dimensional space $\mathbb{R}^N_{\boldsymbol{M}}$ as the Euclidean space $\mathbb{R}^N$ equipped with a weighted inner product $(\boldsymbol{m}_1, \boldsymbol{m}_2)_{\boldsymbol{M}} := \boldsymbol{m}_1^T \boldsymbol{M} \boldsymbol{m}_2$, where $\boldsymbol{m}_1, \boldsymbol{m}_2 \in \mathbb{R}^N$. For functions $u \in V_N$, we have $u = \sum_{k=1}^N u_k \phi_k$ and denote $\boldsymbol{u} := (u_1, \dots, u_N)^T$. Obviously, we have $(m_1, m_2)_{L^2(\Omega)} \approx (\boldsymbol{m}_1, \boldsymbol{m}_2)_{\boldsymbol{M}}$ for $m_1, m_2 \in V_N \subset L^2(\Omega)$. We denote $\boldsymbol{S}$ as the discretized measurement operator that maps from $\mathbb{R}^N_{\boldsymbol{M}}$ to $\mathbb{R}^{N_d}$. Here, N_d is the number of measured points. We denote $\boldsymbol{S}^\natural$ as the adjoint of $\boldsymbol{S}$, which is an operator that maps from $\mathbb{R}^{N_d}$ to $\mathbb{R}^N_{\boldsymbol{M}}$. By the definition of the adjoint operator, it follows that $(\boldsymbol{S}\boldsymbol{u}, \boldsymbol{d}) = (\boldsymbol{u}, \boldsymbol{S}^\natural \boldsymbol{d})_{\boldsymbol{M}}$ for all $\boldsymbol{u} \in \mathbb{R}^N_{\boldsymbol{M}}$ and $\boldsymbol{d} \in \mathbb{R}^{N_d}$, which implies that $\boldsymbol{u}^T \boldsymbol{S}^T \boldsymbol{d} = \boldsymbol{u}^T \boldsymbol{M} \boldsymbol{S}^\natural \boldsymbol{d}$. That is to say, we have $\boldsymbol{S}^\natural = \boldsymbol{M}^{-1} \boldsymbol{S}^T$.

With these illustrations, it becomes evident that we can compute the transformation of the discrete data points $\boldsymbol{d} \in \mathbb{R}^{N_d}$ to a discretized function using $\boldsymbol{M}^{-1}\boldsymbol{S}^T\boldsymbol{d}$, thereby allowing this function to be used as input for the FNO. Due to the special structures of the FNO, its performance is not expected to be significantly affected by the discretization dimensions.

It should be noted that the finite element method is implemented by employing the open-source software FEniCS (version 2019.1.0). For additional information on FEniCS, see Logg et al. (2012). The related neural networks are implemented using the open-source software PyTorch (version 1.11.0+cu113). The FNO are implemented with minor modifications based on the code provided in the original paper (Li et al. 2020). All of the stiffness matrices and mass matrices generated by FEniCS are converted to NumPy arrays and PyTorch tensors, which are utilized to reformulate the forward and adjoint equations within a neural network layer.

Simple environment of the backward diffusion problem. In the main text, we present numerical results under a complex environment setting, which means that the true functions are generated from a distribution with two major branches. As established in the main text, a learning algorithm with a data-independent prior $\mathcal{N}(f(\theta), \mathcal{C}_0)$ cannot learn useful information. However, a learning algorithm with a data-dependent prior $\mathcal{N}(f(S, \theta), \mathcal{C}_0)$ can extract valuable information from historical inverse tasks. In the following, let us provide numerical results for the backward diffusion problem under a simple environment setting, where the true functions are generated from a distribution with only one major branch. As shown in the following illustrations, we will find that the learning algorithm with a data-dependent prior also performs better than the algorithm with a data-independent prior. The illustrations consistently show that only averaged information is learned under the data-independent prior assumption, whereas more detailed information can be extracted under the data-dependent prior assumption.

First, we provide the specific parameter settings for learning $\mathbb{P}^\theta$ and $\mathbb{P}_S^\theta$, which were not detailed in the main text. For learning the parameter θ of $\mathbb{P}^\theta$ under both the simple (shown below) and complex (shown in the main text) environment cases, we set the learning rate to 10^{-2}. For learning the parameter θ of $\mathbb{P}_S^\theta$ under both the simple and complex environment cases, we set the initial learning rate to 10^{-2}, adjust it to 0.5×10^{-3} for iterations between 100 and 1000, and further reduce it to 10^{-3} for iterations beyond 1000. For the simple and complex environment cases, we set the total number of iterations to 2000 and 4000, respectively.

Now, let us introduce the following randomized function

$$u(x) = (5\beta x + a\sin(2(5x - b)) + c)e^{-20\left(x-\frac{1}{2}\right)^2}, \tag{38}$$

where $\beta \sim \mathcal{N}(0.5, 0.5)$, $a \sim U(5, 15)$, $b \sim U(0, 0.1)$, and $c \sim \mathcal{N}(4, 1)$. Here $U(x_1, x_2)$ stands for the uniform distribution with lower and upper bounds of x_1 and x_2. We generate 1000 random functions according to the formula (15). Then we generate the datasets by solving the forward diffusion equations and add noises with a noise level of 0.1, i.e., $\eta \sim \mathcal{N}(0, 0.1^2\mathrm{Id})$. Before giving quantitative comparisons, let us provide a visual comparison in Figure **??**. In Figure **??** (a), we show five random functions generated based on formula (15), which gives an intuitive sense of the model parameters. In part (b) of Figure **??**, we show the background true function by the black line, the learned prior mean function with the data-independent assumption (prior mean function assumed to be $f(\theta)$) by the dashed blue line, and the learned prior mean function with the data-dependent assumption (prior mean function assumed to be $f(S; \theta)$) by the dash-dotted red line. In part (b) of Figure **??**, we observe that both learned prior mean functions reflect the main characteristics of the true function. Furthermore, the learned prior mean function with the data-dependent assumption is visually more similar to the background true function. In part (c) of Figure

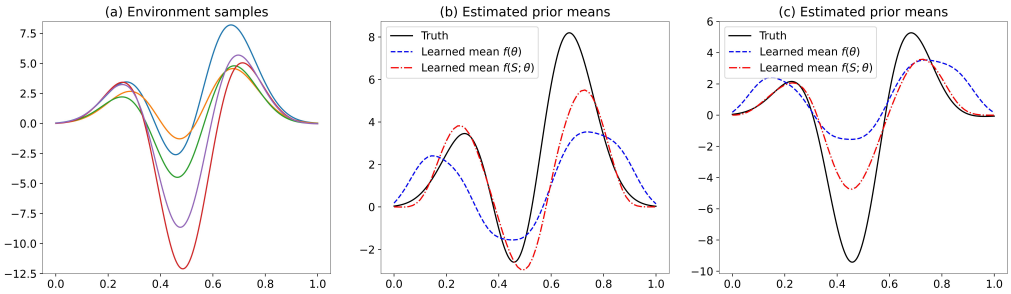

Figure 1: Random functions generated according to formula (15). (a) Ground truth functions used for constructing the datasets. (b)(c) Learned mean functions: the data-independent assumption $f(\theta)$ (dashed blue line) and the data-dependent assumption $f(S;\theta)$ (dash-dotted red line).

??, we provide another background source function and the estimated mean functions obtained from the learned prior means under both data-independent and data-dependent assumptions, which also indicates that the data-dependent learned prior mean can better capture the characteristics of the background true function.

After the visual comparison, we now provide quantitative comparisons in the left part of Table 1. The terms "ItN" and "relative error" are defined as in the main text. The average relative errors displayed in the left part of Table 1 are calculated based on 100 testing examples generated according to (15). It is evident that the learned prior measures converge faster compared to the optimization problem with prior measures that are not learned. Furthermore, under the current setting, the final posterior mean estimates derived from data-independent learned priors exhibit smaller relative errors than those of the unlearned prior case. Additionally, the final posterior mean estimates obtained using data-dependent learned prior measures show significantly smaller relative errors compared to those obtained from the data-independent learned prior case.

Numerical results for the Darcy flow problem. In the main text, we have omitted all details regarding the steady-state Darcy flow problem. Now, let us provide these details. Let $\Omega \subset \mathbb{R}^2$ be a bounded open set, and define the Hilbert space $\mathcal{U}$ as $(L^2(\Omega), \langle\cdot,\cdot\rangle, \|\cdot\|)$. The Darcy flow equation takes the following form:

$$\begin{aligned} -\nabla \cdot (e^u \nabla w) &= f \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{39}$$

where f represents the sources, and $e^{u(x)}$ describes the permeability of the porous medium.

Elements of the inverse problems for Darcy flow:

- Datasets: At a set of discrete points $\{x_j\}_{j=1}^m$, we measure the value of the solution w through a measurement operator $\mathcal{L}_{x_j}$, defined as

$$\mathcal{L}_{x_j}(w) = \int_\Omega \frac{1}{2\pi\delta^2} e^{-\frac{1}{2\delta^2}\|x-x_j\|^2} w(x)dx,$$

  with $\delta > 0$ being a sufficiently small number, and $x_j \in \Omega$ for $j = 1, \ldots, m$.

- Interested parameter: The function $u(x)$.

Table 1: Under simple environment, we assess the average relative errors of maximum a posteriori estimates using unlearned, data-independently learned, and data-dependently learned prior measures. Here, ItN denotes the iteration count of the applied inexact matrix-free Newton-conjugate gradient method.

	Backward diffusion problem				Darcy flow problem		
ItN	$\mathcal{N}(0,\mathcal{C}_0)$	$\mathcal{N}(f(\theta),\mathcal{C}_0)$	$\mathcal{N}(f(S;\theta),\mathcal{C}_0)$	ItN	$\mathcal{N}(0,\mathcal{C}_0)$	$\mathcal{N}(f(\theta),\mathcal{C}_0)$	$\mathcal{N}(f(S;\theta_1),\mathcal{C}_0(\theta_2))$
1	.7458	.2892	**.1901**	1	.7356	.2093	**.0822**
2	.6511	.2435	**.1747**	5	.2999	.1861	**.0737**
3	.6178	.2036	**.1639**	10	.1973	.1860	**.0736**
4	.5810	.1899	**.1609**	15	.1629	.1860	**.0726**
5	.5381	.1833	**.1584**	20	.1414	.1860	**.0710**
6	.5189	.1811	**.1575**	25	.1287	.1860	**.0709**
7	.5138	.1803	**.1575**	30	.1184	.1860	**.0702**

For the Darcy flow problem, we focus on the case where $\Omega = (0,1)^2$. To prevent the inverse crime, we utilize a mesh consisting of 200 grid points to generate the datasets. Additionally, we employ a mesh with 50 grid points to learn the mean function of the base prior measure. Similar to the approach taken in the backward diffusion problem, we set the sampling number $L = 10$ and the mini-batch number $H = 10$.

In Section 3, we apply our general theories to learn prior mean functions. However, these theories could be easily adapted to learn both the prior mean function and the covariance operator. While this simple generalization is possible, it would make the presentation more technical and potentially harder to read, so we choose to present only the theories related to learning prior mean functions. For instance, the covariance operator of the base prior measure $\mathcal{C}_0$ can be decomposed as follows:

$$\mathcal{C}_0 = \sum_{k=1}^{\infty} \lambda_k^2 e_k \otimes e_k.$$

We now fix a positive integer N_c and introduce the following operator:

$$\mathcal{C}_0(\theta_2) = \sum_{k=1}^{N_c} e^{\theta_{2k}} e_k \otimes e_k + \sum_{k=N_c+1}^{\infty} \lambda_k^2 e_k \otimes e_k,$$

where $\theta_2 = (\theta_{21}, \ldots, \theta_{2N_c})$. When $e^{\theta_{2k}} = \lambda_k^2$ for $k = 1, \ldots, N_c$, we have $\mathcal{C}_0(\theta_2) = \mathcal{C}_0$. Thus, the parameters θ_2 can be recognized as learnable parameters. Now, we can define $\theta = (\theta_1, \theta_2)$, and the three choices for the base prior measure $\mathbb{P}$ are specified as follows:

- Unlearned prior measure, i.e., $\mathbb{P} := \mathcal{N}(0, \mathcal{C}_0)$;
- Learned data-independent prior measure, i.e., $\mathbb{P}^\theta := \mathcal{N}(f(\theta), \mathcal{C}_0)$;

- Learned data-dependent prior measure, i.e., $\mathbb{P}_S^\theta := \mathcal{N}(f(S;\theta_1), \mathcal{C}_0(\theta_2))$.

When the mean function of the base prior is assumed to be data-independent, we take the hyperprior $\mathscr{P} := \mathcal{N}(0, A^{-2}) \otimes \mathcal{N}(2\ln \boldsymbol{\lambda}, 10\mathrm{Id}_{N_c})$, where $A = 0.01\mathrm{Id} - 0.1\Delta$, $2\ln \boldsymbol{\lambda} := (2\ln\lambda_1, \ldots, 2\ln\lambda_{N_c})$, and Id_{N_c} is the identity operator on $\mathbb{R}^{N_c}$. When the mean function of the base prior takes the form $f(S;\theta_1)$, we select the hyperprior $\mathscr{P} := g_{\#}\mathcal{N}(0, A^{-2}) \otimes \mathcal{N}(2\ln \boldsymbol{\lambda}, 10\mathrm{Id}_{N_c})$, with g defined as in the backward diffusion problem.

In our numerical illustrations, we select $N_c = 45$ and refer to Bui-Thanh & Nguyen (2016) for detailed strategies on the discretization of the base prior measures. Additionally, other parametric strategies for the covariance operator $\mathcal{C}_0$ can be adopted, as suggested in Pinski et al. (2015*a*), to construct more flexible learning methods. For the Fourier neural operator, we set the maximal number of Fourier modes to 12 for each Fourier integral operator, and the original inputs are mapped to a higher-dimensional representation with a dimension equal to 12. For learning the parameter θ of $\mathbb{P}_S^\theta$ in both simple and complex environments, we set the initial learning rate to 10^{-3}. We then adjust the learning rate to 10^{-4} when the iteration number reaches between 100 and 500. When the iteration number exceeds 500, we further reduce the learning rate to 10^{-5}. In the case of simple and complex environments, we set the total number of iterations to 4000 and 7000, respectively.

Simple environment of the Darcy flow problem. Let us define the following random function:

$$u(x_1, x_2) = u_1(x_1, x_2) + u_2(x_1, x_2) + u_3(x_1, x_2), \tag{40}$$

where

$$u_i(x_1, x_2) = a_{i3}(1 - x_1^2)^{a_{i1}}(1 - x_2^2)^{a_{i2}} e^{-a_{i4}(x_1 - a_{i5})^2 - a_{i6}(x_2 - a_{i7})^2} \tag{41}$$

with $i = 1, 2, 3$. For the parameters in the above formula, we let

$$\begin{aligned}
&a_{i1} \sim U(0.1, 0.5), \quad a_{i2} \sim U(0.1, 0.5), \quad a_{i3} \sim U(3, 4), \quad a_{i4} \sim U(30, 35),\\
&a_{i6} \sim U(30, 35), \quad a_{15} \sim U(0.15, 0.25), \quad a_{17} \sim U(0.65, 0.75), \quad a_{25} \sim U(0.45, 0.55),\\
&a_{27} \sim U(0.45, 0.55), \quad a_{35} \sim U(0.65, 0.75), \quad a_{37} \sim U(0.15, 0.25),
\end{aligned}$$

where $i = 1, 2, 3$. We generate 2000 random functions according to formula (17). As illustrated in Subsection 2.1 of the main text, we select sparse measurement locations equally spaced in a statistical sense, with the number of points set to $m = 100$. By solving the Darcy flow equation, we obtain the dataset by measuring the solution at the measurement points and adding noise with a level of $0.01 \max(|\boldsymbol{y}|)$, where $\boldsymbol{y}$ represents the vector of measured data.

Before presenting quantitative comparisons, we provide a visual comparison in Figure 1. In Figure 1(a), we display one of the random functions, offering an intuitive understanding of the model parameters. In Figure 1(b), we present the estimated mean function obtained by the FNO under the data-dependent prior assumption $\mathbb{P}_S^\theta$. Lastly, in Figure 1(c), we display the estimated mean function under the data-independent prior assumption $\mathbb{P}^\theta$. Notably, the estimated mean function shown in (b) of Figure 1 closely resembles the true underlying function. In contrast, the estimated mean function depicted in (c) of Figure 1 only represents the averaged information and visually deviates from this specific true underlying function.

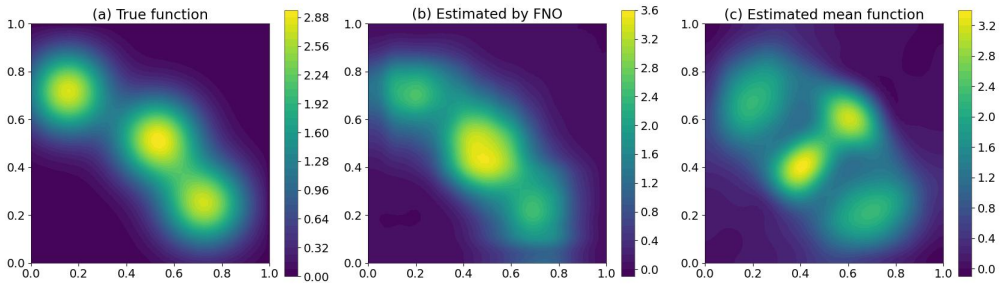

Figure 2: Random functions and estimated mean functions when the datasets are constructed through formula (17). (a) One of the random function generated based on formula (17); (b) The learned mean function by the data-dependent prior mean $f(S;\theta)$; (c) The learned mean function by the data-independent prior mean $f(\theta)$.

Following the visual comparison, we present quantitative comparisons in the right portion of Table 1. The notations used here are consistent with those employed in the backward diffusion problem. The average relative errors displayed in the right portion of Table 1 are computed based on 100 test examples generated according to formula (17). An examination of the right portion of Table 1 reveals that both learned prior measures converge significantly faster than the optimization problem with an unlearned prior measure, particularly during the initial iterations. This indicates that the learned prior effectively captures useful information from related historical inverse tasks.

Since the data-independent learned prior $\mathbb{P}^{\theta}$ can only extract prior information on average, the data-dependent learned prior $\mathbb{P}_S^{\theta}$ demonstrates superior performance. As observed in the right portion of Table 1, the relative errors for the data-dependent learned prior case are approximately 7 to 8 times smaller than those for the data-independent learned prior case. This finding underscores the advantage of incorporating data-dependent learned prior measures, even within a simple environmental setting.

Complex environment of the Darcy flow problem. For the complex environment, the background true functions are generated based on the following formula:

$$u(x_1, x_2) = (2\alpha - 1)(u_1(x_1, x_2) + u_2(x_1, x_2) + u_3(x_1, x_2)), \tag{42}$$

where $\alpha \sim \text{Bern}(0.5)$ and

$$u_i(x_1, x_2) = a_{i3}(1 - x_1^2)^{a_{i1}}(1 - x_2^2)^{a_{i2}} e^{-a_{i4}(x_1 - a_{i5})^2 - a_{i6}(x_2 - a_{i7})^2}$$

with $i = 1, 2, 3$. For the parameters in the above formula, we let

$$\begin{aligned} &a_{i1} \sim U(0.1, 0.5), \quad a_{i2} \sim U(0.1, 0.5), \quad a_{i3} \sim U(3, 4), \quad a_{i4} \sim U(30, 35), \\ &a_{i6} \sim U(30, 35), \quad a_{15} \sim U(0.15, 0.25), \quad a_{17} \sim U(0.65, 0.75), \quad a_{25} \sim U(0.45, 0.55), \\ &a_{27} \sim U(0.45, 0.55), \quad a_{35} \sim U(0.65, 0.75), \quad a_{37} \sim U(0.15, 0.25), \end{aligned}$$

where $i = 1, 2, 3$. We generate 2000 random functions according to formula (19). Subsequently, we create the datasets by solving the Darcy flow equation and adding noise at a level of $0.01 \max(|\boldsymbol{y}|)$, where $\boldsymbol{y}$ represents the vector of measured data. The parameter α introduces two primary branches in the randomized model parameters. As before, we also

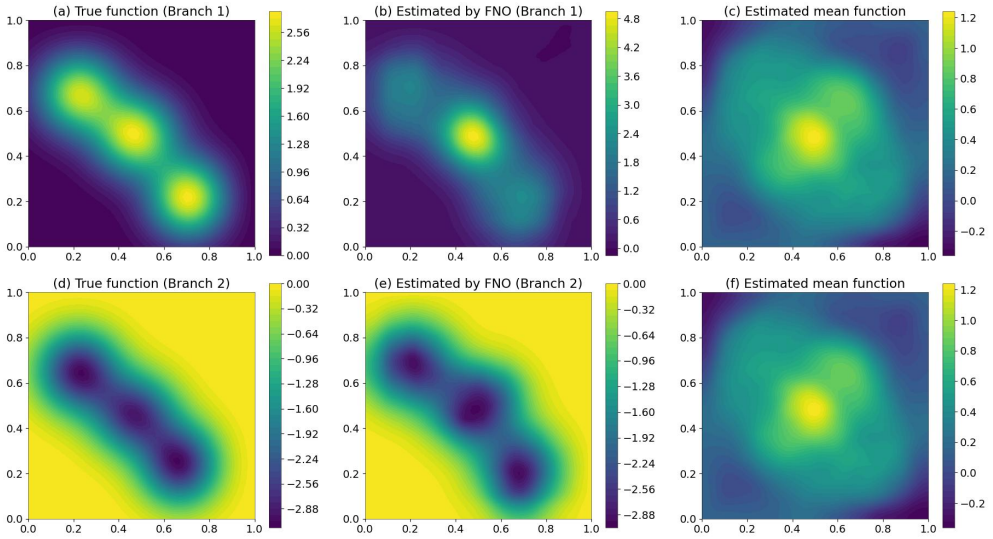

Figure 3: Random and estimated functions for Branches 1 and 2: (a) illustrates a random function of Branch 1; (b) and (c) show the estimated mean functions from data-dependent and data-independent priors for Branch 1, respectively; (d) introduces a random function of Branch 2; while (e) and (f) present the corresponding estimated mean functions from both priors for Branch 2.

generate 100 testing examples. In Figures 2(a) and 2(d), we display two functions randomly selected from the two main branches of the testing examples. In Figures 2(b) and 2(e), we present the learned mean function under the data-dependent prior assumption. Clearly, the learned Fourier neural operator can distinguish between the two main branches and produce reasonable estimates. In Figures 2(c) and 2(f) (which are actually the same figure), we provide the learned mean function under the data-independent prior assumption. From these figures, it is evident that the data-independent learned mean function fails to capture the primary characteristics of the two branches sufficiently.

Finally, we present quantitative comparisons of the maximum a posteriori (MAP) estimations obtained with different priors in the right section of Table 1 of the main text. The average relative errors are computed based on 100 test examples. In the case of the current complex environment, the learned data-independent prior $\mathbb{P}^{\theta}$ yields even less accurate estimates compared to the unlearned prior $\mathbb{P} = \mathcal{N}(0, \mathcal{C}_0)$. This learned prior measure $\mathbb{P}^{\theta}$ fails to provide useful information and may even mislead the base MAP solver. In contrast, the learned data-dependent prior $\mathbb{P}_S^{\theta}$ remains highly informative, enabling the base MAP solver to achieve a high-quality estimate (as measured by the L^2-norm based relative errors) within just 5 iterations. The results presented here clearly demonstrate that the data-dependent prior is capable of extracting valuable information from historical inverse tasks, utilizing only the historical measurement datasets without the true parameter.

B Proof Details

B.1 Proof Details of the Main Text

In this supplementary section, we will offer detailed proofs for all the theorems and lemmas that were mentioned but not elaborated upon in the main text. These proofs are essential for a rigorous understanding of the mathematical framework and the theoretical underpinnings of the concepts discussed.

Proof of Theorem 1 (The General PAC-Bayesian Bound)

Proof. For proving Theorem 1, we need to introduce an intermediate quantity, the expected multi-task error:

$$\tilde{\mathcal{L}}(\mathscr{Q}, \mathbb{D}_1, \ldots, \mathbb{D}_n) = \mathbb{E}_{\theta\sim\mathscr{Q}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathbb{Q}(S_i, \mathbb{P}^{\theta}_{S_i}), \mathbb{D}_i)\right].$$

Let us recall the Donsker-Varadhan variational formula (formula (A.1) in Pinski et al. (2015*b*))

$$D_{\mathrm{KL}}(\mathsf{Q}||\mathsf{P}) = \sup_{g}\left(\mathbb{E}_{\mathsf{Q}}[g] - \ln \mathbb{E}_{\mathsf{P}}[e^{g}]\right), \tag{43}$$

where the measures Q and P are defined on some separable Banach space $\mathcal{W}$, and the supremum are taken all bounded measurable functions $g : \mathcal{W} \to \mathbb{R}$. From the above Donsker-Varadhan variational formula (20), we easily obtain the following inequality

$$\mathbb{E}_{\mathsf{Q}}[g] \leq \frac{1}{\zeta}\Big(D_{\mathrm{KL}}(\mathsf{Q}||\mathsf{P}) + \ln \mathbb{E}_{\mathsf{P}}[e^{\zeta g}]\Big) \tag{44}$$

holds for any positive real number $\zeta > 0$ and any integrable measurable function g.

In the following, we assume that the hyper-posterior measure $\mathscr{Q}$ is absolutely continuous with respect to the hyper-prior measure $\mathscr{P}$. For every $i = 1, \ldots, n$, we assume that the posterior measure $\mathbb{Q}(S_i, \mathbb{P}^{\theta}_{S_i})$ is absolutely continuous with respect to the prior measure $\mathbb{P}^{\theta}_{S_i}$ for all $S_i \in \mathcal{Z}^m$ and $\theta \in \Theta$. If any of these absolute continuity assumptions are violated, the right-hand side of the estimate (3) of the main text will be infinite. Hence the estimate (3) holds true.

Step 1 (Base-posterior learning generalization): The posterior measure $\mathbb{Q}$ depends on the datasets and the prior measure. Given a dataset $S_i \sim \mathbb{D}_i^m$ of size m and a prior probability measure $\mathbb{P}^{\theta}_{S_i}$, we denote $\mathbb{Q}_i = \mathbb{Q}(S_i, \mathbb{P}^{\theta}_{S_i})$ be the posterior measure of the parameter $u_i \in \mathcal{U}$ for $i = 1, \ldots, n$. Let $S' = \cup_{i=1}^{n} S_i$ be the union of all the datasets. Let $\mathcal{W} = \Theta \times \prod_{i=1}^{n}\mathcal{U}$, $\mathsf{P}(d\theta, du_1, \ldots, du_n) := \left(\otimes_{i=1}^{n}\mathbb{P}^{\theta}_{S_i}(du_i)\right)\mathscr{P}(d\theta)$, $\mathsf{Q}(d\theta, du_1, \ldots, du_n) := \left(\otimes_{i=1}^{n}\mathbb{Q}_i(du_i)\right)\mathscr{Q}(d\theta)$, and

$$g(\theta, u_1, \ldots, u_n) := \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\mathbb{E}_{z\sim\mathbb{D}_i}\left[\frac{1}{nm}\ell(u_i, z)\right] - \frac{1}{nm}\ell(u_i, z_{ij})\right).$$

Now, we can use inequality (**??**) directly to derive

$$\mathbb{E}_{\theta\sim\mathscr{Q}}\mathbb{E}_{u_1\sim\mathbb{Q}_1}\cdots\mathbb{E}_{u_n\sim\mathbb{Q}_n}[g] \leq \frac{1}{\gamma}D_{\mathrm{KL}}\left(\mathsf{Q}||\mathsf{P}\right) + \Pi^{1}(\gamma), \tag{45}$$

where

$$\Pi^1(\gamma) = \frac{1}{\gamma} \ln \mathbb{E}_{\theta \sim \mathscr{P}} \mathbb{E}_{u_1 \sim \mathbb{P}^{\theta}_{S_1}} \cdots \mathbb{E}_{u_n \sim \mathbb{P}^{\theta}_{S_n}} \exp\left(\frac{\gamma}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} [\mathbb{E}_{z \sim \mathbb{D}_i} \ell(u_i, z) - \ell(u_i, z_{ij})] \right).$$

For the left-hand side of (**??**), we have

$$\begin{aligned} \mathbb{E}_{\theta \sim \mathscr{Q}} \mathbb{E}_{u_1 \sim \mathbb{Q}_1} \cdots \mathbb{E}_{u_n \sim \mathbb{Q}_n}[g] =& \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta \sim \mathscr{Q}} \mathbb{E}_{u_i \sim \mathbb{Q}_i} \mathcal{L}(u_i, \mathbb{D}_i) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta \sim \mathscr{Q}} \hat{\mathcal{L}}(\mathbb{Q}_i, S_i) \\ =& \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta \sim \mathscr{Q}} \mathcal{L}(\mathbb{Q}_i, \mathbb{D}_i) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta \sim \mathscr{Q}} \hat{\mathcal{L}}(\mathbb{Q}_i, S_i). \end{aligned} \tag{46}$$

For the first term on the right-hand side of (**??**), we have

$$\begin{aligned} D_{\mathrm{KL}}(\mathsf{Q} || \mathsf{P}) =& \mathbb{E}_{\theta \sim \mathscr{Q}} \mathbb{E}_{u_1 \sim \mathbb{Q}_1} \cdots \mathbb{E}_{u_n \sim \mathbb{Q}_n} \ln \left(\frac{d\mathscr{Q}}{d\mathscr{P}}(\theta) \prod_{i=1}^{n} \frac{d\mathbb{Q}_i}{d\mathbb{P}^{\theta}_{S_i}}(u_i) \right) \\ =& \mathbb{E}_{\theta \sim \mathscr{Q}} \ln \frac{d\mathscr{Q}}{d\mathscr{P}} + \sum_{i=1}^{n} \mathbb{E}_{\theta \sim \mathscr{Q}} \mathbb{E}_{u_i \sim \mathbb{Q}_i} \ln \frac{d\mathbb{Q}_i}{d\mathbb{P}^{\theta}_{S_i}} \\ =& D_{\mathrm{KL}}(\mathscr{Q} || \mathscr{P}) + \sum_{i=1}^{n} \mathbb{E}_{\theta \sim \mathscr{Q}} D_{\mathrm{KL}}(\mathbb{Q}_i || \mathbb{P}^{\theta}_{S_i}). \end{aligned} \tag{47}$$

Plugging formulas (**??**) and (**??**) into inequality (**??**), we finally arrive at

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta \sim \mathscr{Q}} \mathcal{L}(\mathbb{Q}_i, \mathbb{D}_i) \leq & \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta \sim \mathscr{Q}} \hat{\mathcal{L}}(\mathbb{Q}_i, S_i) + \frac{1}{\gamma} D_{\mathrm{KL}}(\mathscr{Q} || \mathscr{P}) \\ & + \frac{1}{\gamma} \sum_{i=1}^{n} \mathbb{E}_{\theta \sim \mathscr{Q}} D_{\mathrm{KL}}(\mathbb{Q}_i || \mathbb{P}^{\theta}_{S_i}) + \Pi^1(\gamma). \end{aligned} \tag{48}$$

Step 2 (Hyper-posterior learning generalization): Now, we choose the space $\mathcal{W} = \Theta$, $\mathsf{Q} = \mathscr{Q}$, $\mathsf{P} = \mathscr{P}$, and

$$g(\theta) = \sum_{i=1}^{n} \left(\mathbb{E}_{\mathbb{D} \sim \mathcal{T}} \mathbb{E}_{S \sim \mathbb{D}^m} \left[\frac{1}{n} \mathcal{L}(\mathbb{Q}(S, \mathbb{P}^{\theta}_S), \mathbb{D}) \right] - \frac{1}{n} \mathcal{L}(\mathbb{Q}(S_i, \mathbb{P}^{\theta}_{S_i}), \mathbb{D}_i) \right)$$

in inequality (**??**). Then, we obtain

$$\mathbb{E}_{\theta \sim \mathscr{Q}}[g(\theta)] \leq \frac{1}{\lambda} D_{\mathrm{KL}}(\mathscr{Q} || \mathscr{P}) + \Pi^2(\lambda), \tag{49}$$

where

$$\Pi^2(\lambda) = \frac{1}{\lambda} \ln \mathbb{E}_{\theta \sim \mathscr{P}} \exp\left(\frac{\lambda}{n} \sum_{i=1}^{n} \left[\mathbb{E}_{\mathbb{D} \sim \mathcal{T}} \mathbb{E}_{S \sim \mathbb{D}^m} \left[\mathcal{L}(\mathbb{Q}(S, \mathbb{P}^{\theta}_S), \mathbb{D}) \right] - \mathcal{L}(\mathbb{Q}(S_i, \mathbb{P}^{\theta}_{S_i}), \mathbb{D}_i) \right] \right).$$

By a simple reduction for the left-hand side of (**??**), we obtain

$$\mathcal{L}(\mathscr{Q},\mathcal{T}) \leq \tilde{L}(\mathscr{Q},\mathbb{D}_1,\ldots,\mathbb{D}_n) + \frac{1}{\lambda} D_{\mathrm{KL}}(\mathscr{Q}||\mathscr{P}) + \Pi^2(\lambda). \tag{50}$$

Step 3 (Combination): Combining estimates (**??**) and (**??**), we obtain

$$\begin{aligned}\mathcal{L}(\mathscr{Q},\mathcal{T}) \leq & \hat{\mathcal{L}}(\mathscr{Q},S_1,\ldots,S_n) + \left(\frac{1}{\lambda}+\frac{1}{\gamma}\right) D_{\mathrm{KL}}(\mathscr{Q}||\mathscr{P}) \\ & + \frac{1}{\gamma}\sum_{i=1}^{n} \mathbb{E}_{\theta\sim\mathscr{Q}} D_{\mathrm{KL}}(\mathbb{Q}_i||\mathbb{P}_{S_i}^{\theta}) + \Pi^1(\gamma) + \Pi^2(\lambda).\end{aligned} \tag{51}$$

Taking $\epsilon > 0$ be a positive small number, we have

$$\begin{aligned}\mathbb{P}\left[\Pi^1(\gamma)+\Pi^2(\lambda) > \ln\epsilon\right] =& \mathbb{P}\left[\exp\left(\Pi^1(\gamma)+\Pi^2(\lambda)\right) > \epsilon\right] \\ \leq & \mathbb{P}\left[\exp\left(\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda)\right) > \epsilon^{\sqrt{n}}\right] \\ \leq & \frac{\mathbb{E}\left(\exp(\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda))\right)}{\epsilon^{\sqrt{n}}},\end{aligned} \tag{52}$$

where we employed the Markov's inequality to derive the second inequality. Let

$$\delta = \frac{\mathbb{E}\left(\exp(\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda))\right)}{\epsilon^{\sqrt{n}}}, \tag{53}$$

then we find that

$$\ln\epsilon = \frac{1}{\sqrt{n}}\ln\mathbb{E}\left(\exp(\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda))\right) + \frac{1}{\sqrt{n}}\ln\frac{1}{\delta}. \tag{54}$$

Combining the estimates (**??**), (**??**), and (**??**), we proved that

$$\Pi^1(\gamma)+\Pi^2(\lambda) \leq \frac{1}{\sqrt{n}}\ln\mathbb{E}\left(\exp(\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda))\right) + \frac{1}{\sqrt{n}}\ln\frac{1}{\delta}$$

holds true with probability at least $1-\delta$, which implies the desired conclusion. □

Proof of Corollary 1 (Bound for the Sub-Gaussian Loss [Data-Dependent Prior])

Proof. For the term V_i^1, we have

$$\begin{aligned}V_i^1 =& \mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}_{S_i}^{\theta}} \exp\left(\tilde{\gamma}\left[\mathcal{L}(u_i,\mathbb{D}_i)-\ell(u_i,z_i)\right]\right) \\ \leq & \mathbb{E}_{S_i'\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}_{S_i'}^{\theta}} \Psi(S_i,S_i',u_i,\theta)\exp\left(\tilde{\gamma}[\mathcal{L}(u_i,\mathbb{D}_i)-\ell(u_i,z_i)]\right) \\ \leq & \Psi_E^{1/2}\left(\mathbb{E}_{S_i'\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{S_i\sim\mathbb{D}_{\mathcal{T}}}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}_{S_i'}^{\theta}} \exp\left(2\tilde{\gamma}[\mathcal{L}(u_i,\mathbb{D}_i)-\ell(u_i,z_i)]\right)\right)^{1/2} \\ \leq & \Psi_E^{1/2}\exp\left(\tilde{\gamma}^2 s_{\mathrm{I}}^2\right).\end{aligned} \tag{55}$$

Through some simple calculations as for the bounded loss case, we find that

$$\left(\mathbb{E}e^{2\sqrt{n}\Pi^1(\gamma)}\right)^{1/2} \leq \Psi_E^{\frac{n\sqrt{n}}{2\gamma}}\exp\left(\frac{\gamma s_{\mathrm{I}}^2}{\sqrt{n}m}\right). \tag{56}$$

Combining estimates on the term $\mathbb{E}e^{2\sqrt{n}\Pi^2(\lambda)}$ (not given here since it is similar to the data-independent case), we obtain

$$\mathbb{E}\left[e^{\sqrt{n}\Pi^1(\gamma)+\sqrt{n}\Pi^2(\lambda)}\right] \leq \Psi_E^{\frac{n\sqrt{n}}{2\gamma}} \exp\left(\frac{\gamma s_{\mathrm{I}}^2}{\sqrt{n}m} + \frac{\lambda s_{\mathrm{II}}^2}{2\sqrt{n}}\right). \tag{57}$$

Combining estimate (23) with the results given in Theorem 1 of the main text, we obtain the desired result. □

Proof of Theorem 2 (Bayesian Well-Posedness of Inverse Problems)

Proof. In the following, we only present the proofs when $\mathcal{H}$ is assumed to be an infinite-dimensional separable Hilbert space. When $\mathcal{H} = \mathbb{R}^{N_d}$, the proof is similar.

Step 1 (Well defined Bayes' formula). Let us define $\mathbb{P}_0 := \mathcal{N}(0, \mathcal{C}_0)$. Since $f(S;\theta) \in \mathcal{C}_0^{\frac{1}{2}}\mathcal{U}$, the measure $\mathbb{P}_S^\theta$ is equivalent to the measure $\mathbb{P}_0$. Then we transform the Bayes' formulation as follow

$$\frac{d\mathbb{Q}(S,\mathbb{P}_S^\theta)}{d\mathbb{P}_0} = \frac{1}{Z_m}\exp\left(-\frac{\beta}{m}\sum_{j=1}^m \Phi(u;x_j,y_j)\right)\frac{d\mathbb{P}_S^\theta}{d\mathbb{P}_0}. \tag{58}$$

If formula (24) is well defined, we know that formula (6) of the main text is well defined. Based on Theorem 2.23 in Prato & Zabczyk (2014), we find that

$$\frac{d\mathbb{P}_S^\theta}{d\mathbb{P}_0}(u) = \exp\left(\left\langle \mathcal{C}_0^{-\frac{1}{2}} f(S;\theta), \mathcal{C}_0^{-\frac{1}{2}}(u - f(S;\theta))\right\rangle_{\mathcal{U}} - \frac{1}{2}\left\|\mathcal{C}_0^{-\frac{1}{2}} f(S;\theta)\right\|_{\mathcal{U}}^2\right).$$

Thus, formula (24) can be written as

$$\frac{d\mathbb{Q}(S,\mathbb{P}_S^\theta)}{d\mathbb{P}_0} = \frac{1}{Z_m}\exp\left(-\frac{\beta}{m}\sum_{j=1}^m \Phi(u;x_j,y_j) - \tilde{\Phi}(u;S)\right), \tag{59}$$

where

$$\tilde{\Phi}(u;S) := -\left\langle \mathcal{C}_0^{-\frac{1}{2}} f(S;\theta), \mathcal{C}_0^{-\frac{1}{2}}(u - f(S;\theta))\right\rangle_{\mathcal{U}} + \frac{1}{2}\left\|\mathcal{C}_0^{-\frac{1}{2}} f(S;\theta)\right\|_{\mathcal{U}}^2.$$

Considering the Assumptions 1 of the main text and Lemma 3.3 in Agapiou et al. (2013), we notice that $u \in \mathcal{U}^{1-\tilde{s}}$ $\mathbb{P}_S^\theta$-a.s. when $\tilde{s} \in (s_0, 1]$. According to the theory presented in Section 4.1 of Dashti & Stuart (2017), we need to verify

- (**continuity condition**) the potential function

$$\frac{\beta}{m}\sum_{j=1}^m \Phi(u;x_j,y_j) + \tilde{\Phi}(u;S) \in C(\mathcal{U}^{1-\tilde{s}} \times (\mathcal{H}^{\alpha-s})^m; \mathbb{R}),$$

- (**lower bound condition**) for all $u \in \mathcal{U}^{1-\tilde{s}}$, $y_j \in B_{\mathcal{H}^{\alpha-s}}(r) := \{y \in \mathcal{H} : \|y\|_{\mathcal{H}^{\alpha-s}} < r\}$

$$\frac{\beta}{m}\sum_{j=1}^m \Phi(u;x_j,y_j) + \tilde{\Phi}(u;S) \geq -N(r, \|u\|_{\mathcal{U}^{1-\tilde{s}}}),$$

  where $N : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ is monotonic non-decreasing separately in each argument. In addition, $\exp\left(N(r, \|u\|_{\mathcal{U}^{1-\tilde{s}}})\right) \in L^1_{\mathbb{P}_0}(\mathcal{U}^{1-\tilde{s}}; \mathbb{R})$.

Continuity of $\frac{\beta}{m}\sum_{j=1}^m \Phi(u;x_j,y_j)$. For the summation $\frac{\beta}{m}\sum_{j=1}^m \Phi(u;x_j,y_j)$, it is enough to consider one of its term $\Phi(u;x_j,y_j)$. For the continuity of u, we have

$$\begin{aligned}|\Phi(u_1;x_j,y_j)-\Phi(u_2;x_j,y_j)| =&\frac{1}{2}\left|\|\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_j}\mathcal{G}(u_1)\|_{\mathcal{H}}^2-\|\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_j}\mathcal{G}(u_2)\|_{\mathcal{H}}^2\right|\\ &+\left|\left\langle\Gamma^{-\frac{1}{2}}y_j,\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_j}\mathcal{G}(u_1)-\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_j}\mathcal{G}(u_2)\right\rangle_{\mathcal{H}}\right| \\ =&\mathrm{I}_1+\mathrm{I}_2\end{aligned} \tag{60}$$

For the first term on the right-hand side of (25), we have

$$\begin{aligned}\mathrm{I}_1\leq&\frac{\tilde{C}^2}{2}(M_1(\|u_1\|_{\mathcal{U}^{1-\tilde{s}}})+M_1(\|u_2\|_{\mathcal{U}^{1-\tilde{s}}}))\|\mathcal{C}_1^{-\frac{s}{2}}\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_j}(\mathcal{G}(u_1)-\mathcal{G}(u_2))\|_{\mathcal{H}}\\ \leq&\frac{\tilde{C}^2}{2}(M_1(\|u_1\|_{\mathcal{U}^{1-\tilde{s}}})+M_1(\|u_2\|_{\mathcal{U}^{1-\tilde{s}}}))M_2(\|u_1\|_{\mathcal{U}^{1-\tilde{s}}},\|u_2\|_{\mathcal{U}^{1-\tilde{s}}})\|u_1-u_2\|_{\mathcal{U}^{1-\tilde{s}}},\end{aligned} \tag{61}$$

where $\tilde{C}:=\|\mathcal{C}_1^{\frac{s}{2}}\|_{\mathcal{B}(\mathcal{H})}$ is the operator norm of $\mathcal{C}_1^{\frac{s}{2}}$. For the second term on the right-hand side of (25), we have

$$\begin{aligned}\mathrm{I}_2\leq&\|y_j\|_{\mathcal{H}^{\alpha-s}}\|\mathcal{C}_1^{-\frac{s}{2}}\Gamma^{-\frac{1}{2}}(\mathcal{L}_{x_j}\mathcal{G}(u_1)-\mathcal{L}_{x_j}\mathcal{G}(u_2))\|_{\mathcal{H}}\\ \leq&\|y_j\|_{\mathcal{H}^{\alpha-s}}M_2(\|u_1\|_{\mathcal{U}^{1-\tilde{s}}},\|u_2\|_{\mathcal{U}^{1-\tilde{s}}})\|u_1-u_2\|_{\mathcal{U}^{1-\tilde{s}}}.\end{aligned} \tag{62}$$

From the estimates (25), (**??**), and (**??**), we obtain the continuity of Φ for the parameter u. For the variable y_j, we have

$$\begin{aligned}|\Phi(u;x_j,y_j)-\Phi(u;x_j,y_j')| =&\left|\left\langle\Gamma^{-\frac{1}{2}}(y_j-y_j'),\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_j}\mathcal{G}(u)\right\rangle_{\mathcal{H}}\right|\\ \leq&\|y_j-y_j'\|_{\mathcal{H}^{\alpha-s}}M_1(\|u\|_{\mathcal{U}^{1-\tilde{s}}}),\end{aligned}$$

which implies the continuity of Φ for the variable y_j. Hence, we proved

$$\frac{\beta}{m}\sum_{j=1}^m\Phi(u;x_j,y_j)\in C(\mathcal{U}^{1-\tilde{s}}\times(\mathcal{H}^{\alpha-s})^m;\mathbb{R}). \tag{63}$$

Continuity of $\tilde{\Phi}(u;S)$. For the continuity of u, we have

$$\begin{aligned}|\tilde{\Phi}(u_1,S)-\tilde{\Phi}(u_2,S)|\leq&\left|\left\langle\mathcal{C}_0^{-\frac{1}{2}}f(S;\theta),\mathcal{C}_0^{-\frac{1}{2}}(u_1-u_2)\right\rangle_{\mathcal{U}}\right|\\ \leq&\|f(S;\theta)\|_{\mathcal{U}^{1+\tilde{s}}}\|u_1-u_2\|_{\mathcal{U}^{1-\tilde{s}}}.\end{aligned} \tag{64}$$

Let $S=\{(x_j,y_j)\}_{j=1}^m$ and $S'=\{(x_j,y_j')\}_{j=1}^m$, we have

$$\begin{aligned}|\tilde{\Phi}(u,S)-\tilde{\Phi}(u,S')| =&\left|\left\langle\mathcal{C}_0^{-\frac{1}{2}}(f(S;\theta)-f(S';\theta)),\mathcal{C}_0^{-\frac{1}{2}}u\right\rangle_{\mathcal{U}}\right|\\ &+\frac{3}{2}\left|\|\mathcal{C}_0^{-\frac{1}{2}}f(S';\theta)\|_{\mathcal{U}}^2-\|\mathcal{C}_0^{-\frac{1}{2}}f(S;\theta)\|_{\mathcal{U}}^2\right|\\ \leq&\|f(S;\theta)-f(S';\theta)\|_{\mathcal{U}^{1+\tilde{s}}}\|u\|_{\mathcal{U}^{1-\tilde{s}}}\\ &+\frac{3}{2}\left(\|f(S;\theta)\|_{\mathcal{U}^1}+\|f(S';\theta)\|_{\mathcal{U}^1}\right)\|f(S;\theta)-f(S';\theta)\|_{\mathcal{U}^1}.\end{aligned} \tag{65}$$

Combining the estimates (26) and (27), and the assumption (5) of the main text, we obtain $\tilde{\Phi}(u;S) \in C(\mathcal{U}^{1-\tilde{s}} \times (\mathcal{H}^{\alpha-s})^m; \mathbb{R})$.

To verify the lower bound condition, we have

$$\begin{aligned}
\frac{\beta}{m}\sum_{j=1}^{m}\Phi(u;x_j,y_j) &\geq -\frac{\beta}{m}\sum_{j=1}^{m}\|\mathcal{C}_1^{\frac{s}{2}}\Gamma^{-\frac{1}{2}}y_j\|_{\mathcal{H}}\|\mathcal{C}_1^{-\frac{s}{2}}\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_j}\mathcal{G}(u)\|_{\mathcal{H}} \\
&\geq -\frac{\beta}{m}\sum_{j=1}^{m}\|y_j\|_{\mathcal{H}^{\alpha-s}}M_1(\|u\|_{\mathcal{U}^{1-\tilde{s}}}) \\
&\geq -\frac{\beta}{m}\sum_{j=1}^{m}\frac{1}{4\delta_1}\|y_j\|^2_{\mathcal{H}^{\alpha-s}} - \delta_1\beta M_1(\|u\|_{\mathcal{U}^{1-\tilde{s}}})^2,
\end{aligned} \tag{66}$$

and

$$\begin{aligned}
\tilde{\Phi}(u,S) &= -\left\langle \mathcal{C}_0^{-\frac{1}{2}}f(S;\theta), \mathcal{C}_0^{-\frac{1}{2}}u\right\rangle_{\mathcal{U}} + \frac{3}{2}\|\mathcal{C}_0^{-\frac{1}{2}}f(S;\theta)\|^2_{\mathcal{U}} \\
&\geq -\|f(S;\theta)\|_{\mathcal{U}^{1+\tilde{s}}}\|u\|_{\mathcal{U}^{1-\tilde{s}}} + \frac{3}{2}\|\mathcal{C}_0^{-\frac{1}{2}}f(S;\theta)\|^2_{\mathcal{U}} \\
&\geq -\delta_2\|u\|^2_{\mathcal{U}^{1-\tilde{s}}} - \left(\frac{1}{4\delta_2} - \frac{3}{2}\right)\|f(S;\theta)\|^2_{\mathcal{U}^{1+\tilde{s}}}.
\end{aligned} \tag{67}$$

According to the assumptions on $M_1(\cdot)$ and Proposition 1.13 in Prato (2006), we know that

$$\mathbb{E}_{u\sim\mathbb{P}_S^\theta}\exp\left(\delta_1\beta M_1(\|u\|_{\mathcal{U}^{1-\tilde{s}}})^2 + \delta_2\|u\|^2_{\mathcal{U}^{1-\tilde{s}}}\right) < +\infty, \tag{68}$$

when δ_1 and δ_2 are chosen to be small enough positive numbers. Considering the above inequality (30) and $\|f(S;\theta)\|_{\mathcal{U}^{1+\tilde{s}}} \leq M(r)$ if $\|y_j\|_{\mathcal{H}^{\beta-s}} \leq r$ $(j=1,\ldots,m)$, we easily find that the lower bound condition is fulfilled.

Step 2 (Stability with respect to datasets). From the definition of Hellinger distance, we have

$$d_{\text{Hell}}(\mathbb{Q}(S,\mathbb{P}_S^\theta),\mathbb{Q}(S',\mathbb{P}_{S'}^\theta))^2 \leq I_1 + I_2,$$

where

$$\begin{aligned}
I_1 =& \frac{1}{Z_m}\int_{\mathcal{U}^{1-s}}\left[\exp\left(-\frac{\beta}{2m}\sum_{j=1}^{m}\Phi(u;x_j,y_j) - \frac{1}{2}\tilde{\Phi}(u;S)\right)\right. \\
&\left. - \exp\left(-\frac{\beta}{2m}\sum_{j=1}^{m}\Phi(u;x'_j,y'_j) - \frac{1}{2}\tilde{\Phi}(u;S')\right)\right]^2\mathbb{P}_0(du),
\end{aligned}$$

and

$$I_2 = \left|Z_m^{-\frac{1}{2}} - (Z'_m)^{-\frac{1}{2}}\right|^2\int_{\mathcal{U}^{1-\tilde{s}}}\exp\left(-\frac{\beta}{m}\sum_{j=1}^{m}\Phi(u;x'_j,y'_j) - \tilde{\Phi}(u;S')\right)\mathbb{P}_0(du),$$

where Z'_m is the normalization constant concerned with the datasets S'. For the term I_1, we have

$$I_1 \leq \frac{1}{2Z_m}\int_{\mathcal{U}^{1-\tilde{s}}} e^{N(r,\|u\|_{\mathcal{U}^{1-\tilde{s}}})}(I_{11}+I_{12})\mathbb{P}_0(du),$$

where

$$I_{11} = \left| \frac{\beta}{m} \sum_{j=1}^{m} \left[\Phi(u; x_j, y_j) - \Phi(u; x'_j, y'_j) \right] \right|^2,$$
$$I_{12} = \left| \tilde{\Phi}(u; S) - \tilde{\Phi}(u; S') \right|^2.$$

For the term I_{11}, we firstly consider the following calculations

$$\begin{aligned}\Phi(u; x_j, y_j) - \Phi(u; x'_j, y'_j) = &\frac{1}{2}\|\Gamma^{-\frac{1}{2}} \mathcal{L}_{x_j} \mathcal{G}(u)\|^2_{\mathcal{H}} - \frac{1}{2}\|\Gamma^{-\frac{1}{2}} \mathcal{L}_{x'_j} \mathcal{G}(u)\|^2_{\mathcal{H}} \\ &+ \left\langle \Gamma^{-\frac{1}{2}} y'_j, \Gamma^{-\frac{1}{2}} \mathcal{L}_{x'_j} \mathcal{G}(u) \right\rangle_{\mathcal{H}} - \left\langle \Gamma^{-\frac{1}{2}} y'_j, \Gamma^{-\frac{1}{2}} \mathcal{L}_{x_j} \mathcal{G}(u) \right\rangle_{\mathcal{H}} \\ &+ \left\langle \Gamma^{-\frac{1}{2}} y'_j, \Gamma^{-\frac{1}{2}} \mathcal{L}_{x_j} \mathcal{G}(u) \right\rangle_{\mathcal{H}} - \left\langle \Gamma^{-\frac{1}{2}} y_j, \Gamma^{-\frac{1}{2}} \mathcal{L}_{x_j} \mathcal{G}(u) \right\rangle_{\mathcal{H}},\end{aligned}$$

which implies that

$$\begin{aligned}|\Phi(u; x_j, y_j) - \Phi(u; x'_j, y'_j)| \leq &\tilde{C}^2 M_1(\|u\|_{\mathcal{U}^{1-\tilde{s}}}) \|\mathcal{C}_1^{-\frac{s}{2}} \Gamma^{-\frac{1}{2}} (\mathcal{L}_{x_j} \mathcal{G}(u) - \mathcal{L}_{x'_j} \mathcal{G}(u))\|_{\mathcal{H}} \\ &+ M_1(\|u\|_{\mathcal{U}^{1-\tilde{s}}}) \|y_j - y'_j\|_{\mathcal{H}^{\alpha-s}} + \|y'_j\|_{\mathcal{H}^{\alpha-s}} \|\mathcal{C}_1^{-\frac{s}{2}} \Gamma^{-\frac{1}{2}} (\mathcal{L}_{x_j} \mathcal{G}(u) - \mathcal{L}_{x'_j} \mathcal{G}(u))\|_{\mathcal{H}}.\end{aligned} \tag{69}$$

The above estimate (31) yields the following result

$$\mathrm{I}_{11} \to 0, \quad \text{as } x_j \to x'_j \text{ in } \mathcal{X} \text{ and } y_j \to y'_j \text{ in } \mathcal{H}^{\alpha-s}. \tag{70}$$

Concerned with the term I_{12}, estimate (27) already implies

$$\mathrm{I}_{12} \to 0, \quad \text{as } x_j \to x'_j \text{ in } \mathcal{X} \text{ and } y_j \to y'_j \text{ in } \mathcal{H}^{\alpha-s}. \tag{71}$$

Combining the above statements (**??**) and (**??**), we finally obtain

$$\mathrm{I}_1 \to 0, \quad \text{as } x_j \to x'_j \text{ in } \mathcal{X} \text{ and } y_j \to y'_j \text{ in } \mathcal{H}^{\alpha-s} \tag{72}$$

by applying the Lebesgue's dominated convergence theorem. For the term I_2, we have

$$\mathrm{I}_2 \leq C \max\left(Z_m^{-3}, (Z'_m)^{-3}\right) |Z_m - Z'_m|^2.$$

Similar to the derivation of (**??**), we can prove $Z_m \to Z'_m$ as $x_j \to x'_j$ in $\mathcal{X}$ and $y_j \to y'_j$ in $\mathcal{H}^{\alpha-s}$. Hence, we easily obtain

$$\mathrm{I}_2 \to 0, \quad \text{as } x_j \to x'_j \text{ in } \mathcal{X} \text{ and } y_j \to y'_j \text{ in } \mathcal{H}^{\alpha-s}. \tag{73}$$

□

Proof of Lemma 1 (Estimate Sub-Gaussian Parameters for the General Linear Inverse Problem)

Before delving into the details of the proof, let us restate Lemma 1, incorporating the explicit estimations of s_I and s_{II} that were omitted in the main text.

Lemma 3.7. *Assume that Assumptions 1 and 2 are satisfied, and the conditions for the linear forward operator are reformulated as follows:*

$$\|\mathcal{C}_1^{-\frac{\tilde{s}}{2}}\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}u\|_{\mathcal{H}} \leq M_1\|u\|_{\mathcal{U}^{1-\tilde{s}}} \text{ and } \|\mathcal{C}_1^{-\frac{\tilde{s}}{2}}\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u_1-u_2)\|_{\mathcal{H}} \leq M_2\|u_1-u_2\|_{\mathcal{U}^{1-\tilde{s}}}, \quad (74)$$

where M_1 and M_2 are two positive constants. Also, assume that the probability measure $\mathscr{E}$ has compact support, satisfying $supp\,\mathscr{E} \subset B_{\tilde{s}}(R_u) := \{u \in \mathcal{U} : \|u\|_{\mathcal{U}^{1-\tilde{s}}} \leq R_u\}$ with $R_u \in \mathbb{R}^+$. Let $\tilde{C} = \|\mathcal{C}_1^{\frac{\tilde{s}}{2}}\|_{\mathcal{B}(\mathcal{H})}$ and assume that $\tilde{\gamma}\tilde{C}^2M_2^2 \leq \lambda_1^{-1}$, where λ_1 is as defined in Assumptions 1. Furthermore, the base posterior measure is derived from formula (7). Consequently, for s_I^2 and s_{II}^2 as defined in Assumptions 2 and 3, we have

$$\begin{aligned}
s_I^2 =& \frac{8\tilde{C}^2M_2^2R_u^2}{\tilde{\gamma}} - \frac{\ln\det(Id - 2\tilde{\gamma}\tilde{C}^2M_2^2\mathcal{C}_0)}{2\tilde{\gamma}^2} + \frac{\tilde{C}^4M_1^4}{4} + M_1^2\,Tr(\mathcal{C}_1^s),\\
s_{II}^2 =& \frac{\tilde{C}^4M_1^4R_u^4}{4} + \tilde{C}^2M_2^2\frac{12R_u^2 + \frac{m}{\beta}\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}_1^m}\,Tr(\tilde{\mathcal{C}}_p) + \frac{3\beta^2}{m^2}\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}_1^m}\,Tr(K_{\boldsymbol{x}}\Gamma_mK_{\boldsymbol{x}}^*)}{\tilde{\lambda}} +\\
&+ \frac{12\beta^2}{\tilde{\lambda}m^2}R_u^2\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}_1^m}\|\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*(\Gamma_m + \frac{\beta}{m}\mathcal{L}_{\boldsymbol{x}}\mathcal{G}\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*)^{-1}\mathcal{L}_{\boldsymbol{x}}\mathcal{G}\|^2_{\mathcal{B}(\mathcal{U}^{1-\tilde{s}})},
\end{aligned}$$

where $\tilde{\mathcal{C}}_p^{-1} = (\mathcal{L}_x\mathcal{G})^*\Gamma_m^{-1}\mathcal{L}_x\mathcal{G} + \frac{m}{\beta}\mathcal{C}_0^{-1}$ and $K_{\boldsymbol{x}} = \mathcal{C}_0^{\frac{1+\tilde{s}}{2}}(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*(\Gamma_m + \frac{\beta}{m}\mathcal{L}_{\boldsymbol{x}}\mathcal{G}\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*)^{-1}$.

The variance factors s_{I}^2 and s_{II}^2 presented in the Lemma 1 differ slightly from those in the conventional sub-Gaussian case, as they depend on $\tilde{\gamma}$ and $\tilde{\lambda}$. To construct the general estimates, we shall set the parameters to $\tilde{\gamma} = \frac{\gamma}{nm}$ and $\tilde{\lambda} = \frac{\lambda}{n}$. This implies that we must select γ and λ to scale as nm and n, respectively, to ensure that s_{I} and s_{II} are finite.

Proof. **Estimates of V_i^1**: For the measurement data y, we assume that $y = \mathcal{L}_x\mathcal{G}u^\dagger + \eta$ with $u^\dagger$ be the background true parameter sampled from $\mathscr{E}$. Then the loss function can be reformulated as follow:

$$\begin{aligned}
\ell(u,z) =& \frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u)\|_{\mathcal{H}}^2 - \langle\Gamma^{-\frac{1}{2}}y, \Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u)\rangle_{\mathcal{H}}\\
=& \frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x(\mathcal{G}(u) - \mathcal{G}(u^\dagger))\|_{\mathcal{H}}^2 - \frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u^\dagger)\|_{\mathcal{H}}^2\\
&- \langle\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u), \Gamma^{-\frac{1}{2}}\eta\rangle_{\mathcal{H}}.
\end{aligned} \quad (75)$$

Relying on the above reformulation (33), for $i = 1, \ldots, n$, we have

$$\begin{aligned}
\mathcal{L}(u_i, \mathbb{D}_i) - \ell(u_i, z_i) =& \mathbb{E}_{z\sim\mathbb{D}_i}\ell(u_i, z) - \ell(u_i, z_i)\\
=& I_1 + I_2 + I_3,
\end{aligned} \quad (76)$$

where

$$\begin{aligned}
I_1 =& \frac{1}{2}\mathbb{E}_{x\sim\mathbb{D}_1}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x(\mathcal{G}(u) - \mathcal{G}(u^\dagger))\|_{\mathcal{H}}^2 - \frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_i}(\mathcal{G}(u_i) - \mathcal{G}(u^\dagger))\|_{\mathcal{H}}^2,\\
I_2 =& \frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_i}\mathcal{G}(u_i^\dagger)\|_{\mathcal{H}}^2 - \frac{1}{2}\mathbb{E}_{x\sim\mathbb{D}_1}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u_i^\dagger)\|_{\mathcal{H}}^2,\\
I_3 =& \langle\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_i}\mathcal{G}(u_i), \Gamma^{-\frac{1}{2}}\eta_i\rangle_{\mathcal{H}}.
\end{aligned}$$

Let us denote the constant $\tilde{C} := \|\mathcal{C}_1^{\frac{s}{2}}\|_{\mathcal{B}(\mathcal{H})}$. For the term I_1, we have

$$I_1 \leq \frac{\tilde{C}^2}{2}\mathbb{E}_{x\sim\mathbb{D}_1}\|\mathcal{C}_1^{-\frac{s}{2}}\Gamma^{-\frac{1}{2}}\mathcal{L}_x(\mathcal{G}u_i - \mathcal{G}u_i^\dagger)\|_{\mathcal{H}}^2 \leq \frac{\tilde{C}^2 M_2^2}{2}\|u_i - u_i^\dagger\|_{\mathcal{U}^{1-\tilde{s}}}^2.$$

Hence, we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathbb{D}_i\sim\mathcal{T}}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}_{S'}^\theta}\exp\left(\tilde{\gamma}I_1\right)\\
\leq&\mathbb{E}_{\mathbb{D}_i\sim\mathcal{T}}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}_{S'}^\theta}\exp\left(\tilde{\gamma}\tilde{C}^2M_2^2(\|u_i - f(S';\theta)\|_{\mathcal{U}^{1-\tilde{s}}}^2 + \|f(S';\theta) - u_i^\dagger\|_{\mathcal{U}^{1-\tilde{s}}}^2)\right)\\
\leq&\exp\left(\tilde{\gamma}4\tilde{C}^2M_2^2R_u^2\right)\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}_{S'}^\theta}\exp\left(\tilde{\gamma}\tilde{C}^2M_2^2(\|u_i - f(S';\theta)\|_{\mathcal{U}^{1-\tilde{s}}}^2\right)\\
\leq&\exp\left(\tilde{\gamma}4\tilde{C}^2M_2^2R_u^2\right)\mathbb{E}_{u_i\sim\mathbb{P}_0}\exp\left(\tilde{\gamma}\tilde{C}^2M_2^2\|u_i\|_{\mathcal{U}^{1-\tilde{s}}}^2\right),
\end{aligned} \tag{77}$$

where $\mathbb{P}_0 := \mathcal{N}(0, \mathcal{C}_0)$. For the last term of the above inequality (**??**), we have

$$\mathbb{E}_{u_i\sim\mathbb{P}_0}\exp\left(\tilde{\gamma}\tilde{C}^2M_2^2\|u_i\|_{\mathcal{U}^{1-\tilde{s}}}^2\right) \leq \det(\mathrm{Id} - 2\tilde{\gamma}\tilde{C}^2M_2^2\mathcal{C}_0^2)^{-1/2}, \tag{78}$$

where the assumption $\tilde{\gamma}\tilde{C}^2M_2^2 \leq \lambda_1^{-1}$ and Theorem 2.17 in Prato & Zabczyk (2014) are employed. Combining (**??**) and (**??**), we find that

$$\mathbb{E}_{\mathbb{D}_i\sim\mathcal{T}}\mathbb{E}_{\theta\sim\mathscr{P}}\mathbb{E}_{u_i\sim\mathbb{P}_{S'}^\theta}e^{\tilde{\gamma}I_1} \leq \exp\left(\frac{\tilde{\gamma}^2}{2}\left[\frac{8\tilde{C}^2M_2^2R_u^2}{\tilde{\gamma}} - \frac{\ln\det(\mathrm{Id} - 2\tilde{\gamma}\tilde{C}^2M_2^2\mathcal{C}_0)}{2\tilde{\gamma}^2}\right]\right). \tag{79}$$

For the term I_2, we obviously have $\mathbb{E}_{z_i\sim\mathbb{D}_i}I_2 = 0$. According to Assumptions 2 and the compactly supported condition illustrated in Lemma 1 of $\mathscr{E}$ in the main text, we find that

$$\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u_i^\dagger)\|_{\mathcal{H}}^2 \leq \tilde{C}^2\|\mathcal{C}_1^{-\frac{s}{2}}\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u_i^\dagger)\|_{\mathcal{H}}^2 \leq \tilde{C}^2M_1^2. \tag{80}$$

From the above estimates (**??**), we arrive at

$$-\frac{1}{2}\tilde{C}^2M_1^2 \leq \mathrm{I}_2 \leq \frac{1}{2}\tilde{C}^2M_1^2. \tag{81}$$

Now, using the Hoeffding's lemma and estimates (**??**), we find that

$$\mathbb{E}_{\mathbb{D}_i\sim\mathcal{T}}\mathbb{E}_{x_i\sim\mathbb{D}_{i1}}e^{\tilde{\gamma}\mathrm{I}_2} \leq \exp\left(\frac{\tilde{\gamma}^2\tilde{C}^4M_1^4}{8}\right). \tag{82}$$

Let us denote $\{\zeta_k, \varphi_k\}_{k=1}^\infty$ be an eigen-system of the operator $\mathcal{C}_1$. Since $\mathcal{C}_1\Gamma = \Gamma\mathcal{C}_1$, we know that the operator Γ has an eigen-system $\{\gamma_k, \varphi_k\}_{k=1}^\infty$, which has the same eigenbasis as the operator $\mathcal{C}_1$ (Subsection 2.4 in Prato (2006)). Without loss of generality, the eigenvalues $\{\zeta_k\}_{k=1}^\infty$ and $\{\gamma_k\}_{k=1}^\infty$ are all rearranged in a descending order. With these notations, we

have

$$\begin{aligned}
\mathbb{E}_{\eta_i\sim\mathbb{D}_3} e^{\tilde{\gamma} I_3} &= \mathbb{E}_{\eta_i\sim\mathbb{D}_3} \exp\left(\tilde{\gamma}\langle \mathcal{C}_1^{-\frac{s}{2}}\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_i}\mathcal{G}(u_i), \mathcal{C}_1^{\frac{s}{2}}\Gamma^{-\frac{1}{2}}\eta_i\rangle_{\mathcal{H}}\right) \\
&= \mathbb{E}_{\eta_i\sim\mathbb{D}_3} \exp\left(\tilde{\gamma}\sum_{k=1}^{\infty}\langle \mathcal{C}_1^{-\frac{s}{2}}\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_i}\mathcal{G}(u_i), \varphi_k\rangle_{\mathcal{H}}\langle \mathcal{C}_1^{\frac{s}{2}}\Gamma^{-\frac{1}{2}}\eta_i, \varphi_k\rangle_{\mathcal{H}}\right) \\
&\leq \mathbb{E}_{\eta_i\sim\mathbb{D}_3} \prod_{k=1}^{\infty}\exp\left(\tilde{\gamma} M_1 \zeta_k^{\frac{s}{2}}\gamma_k^{-\frac{1}{2}}\langle \eta_i, \varphi_k\rangle_{\mathcal{H}}\right) \\
&= \prod_{k=1}^{\infty}\frac{1}{\sqrt{2\pi}\gamma_k}\int_{\mathbb{R}}\exp\left(\tilde{\gamma} M_1 \zeta_k^{\frac{s}{2}}\gamma_k^{-\frac{1}{2}}\eta_{ik} - \frac{1}{2\gamma_k}\eta_{ik}^2\right) d\eta_{ik} \\
&= \prod_{k=1}^{\infty}\exp\left(\frac{\tilde{\gamma}^2 M_1^2}{2}\zeta_k^s\right) \\
&= \exp\left(\frac{\tilde{\gamma}^2}{2}M_1^2\mathrm{Tr}(\mathcal{C}_1^s)\right),
\end{aligned} \tag{83}$$

where $\eta_{ik} = \langle \eta_i, \varphi_k\rangle_{\mathcal{H}}$. Combining estimates (**??**) and (**??**), we finally obtain the desired inequality

$$V_i^1 \leq \exp\left(\frac{\tilde{\gamma}^2}{2}\left[\frac{8\tilde{C}^2 M_2^2 R_u^2}{\tilde{\gamma}} - \frac{\ln\det(\mathrm{Id} - 2\tilde{\gamma}\tilde{C}^2 M_2^2\mathcal{C}_0)}{2\tilde{\gamma}^2} + \frac{\tilde{C}^4 M_1^4}{4} + M_1^2\mathrm{Tr}(\mathcal{C}_1^s)\right]\right).$$

Estimates of V_i^2: As in the estimates of V_i^1, we assume that the data y_i is generated from some true parameter $u_i^\dagger$ sampled from the measure $\mathscr{E}$. Hence, we have

$$\begin{aligned}
\mathcal{L}(\mathbb{Q}(S_i, \mathbb{P}_{S_i}^\theta), \mathbb{D}_i) &= \mathbb{E}_{u\sim\mathbb{Q}(S_i,\mathbb{P}_{S_i}^\theta)}\mathbb{E}_{z\sim\mathbb{D}_i}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}u\|_{\mathcal{H}}^2 - \langle\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}u, \Gamma^{-\frac{1}{2}}y_i\rangle_{\mathcal{H}}\right] \\
&= \mathbb{E}_{u\sim\mathbb{Q}(S_i,\mathbb{P}_{S_i}^\theta)}\mathbb{E}_{z\sim\mathbb{D}_i}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}u\|_{\mathcal{H}}^2 - \langle\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}u, \Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}u_i^\dagger\rangle_{\mathcal{H}}\right. \\
&\quad \left. - \langle\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}u, \Gamma^{-\frac{1}{2}}\boldsymbol{\eta}\rangle_{\mathcal{H}}\right] \\
&= \mathbb{E}_{u\sim\mathbb{Q}(S_i,\mathbb{P}_{S_i}^\theta)}\mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u - u_i^\dagger)\|_{\mathcal{H}}^2\right] \\
&\quad - \mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}u_i^\dagger\|_{\mathcal{H}}^2\right],
\end{aligned} \tag{84}$$

which yields

$$V_i^2 \leq \mathbb{E}_{\mathbb{D}_i\sim\mathcal{T}}\mathbb{E}_{S_i\sim\mathbb{D}_i^m}\mathbb{E}_{\theta\sim\mathscr{P}}\exp\left(\tilde{\lambda}\mathrm{I}_1 + \tilde{\lambda}\mathrm{I}_2\right), \tag{85}$$

where

$$\begin{aligned}
\mathrm{I}_1 &= \mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathbb{E}_{u\sim\mathbb{Q}(S,\mathbb{P}_S^\theta)}\mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u - u^\dagger)\|_{\mathcal{H}}^2\right], \\
\mathrm{I}_2 &= \mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}u_i^\dagger\|_{\mathcal{H}}^2\right] - \mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}u^\dagger\|_{\mathcal{H}}^2\right].
\end{aligned}$$

For the term I_2, we easily obtain

$$-\frac{1}{2}\tilde{C}^2 M_1^2 R_u^2 \leq \mathrm{I}_2 \leq \frac{1}{2}\tilde{C}^2 M_1^2 R_u^2.$$

The above boundedness estimate combined with the Hoeffding's lemma yields

$$\mathbb{E}_{\mathbb{D}_i\sim\mathcal{T}}\mathbb{E}_{S_i\sim\mathbb{D}_i^m} e^{\tilde{\lambda}\mathrm{I}_2} \leq \exp\left(\frac{\tilde{\lambda}^2\tilde{C}^4 M_1^4 R_u^4}{8}\right). \tag{86}$$

For the term I_1, we have

$$\begin{aligned}
\mathrm{I}_1 \leq& \frac{1}{2}\tilde{C}^2 M_2^2 \mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathbb{E}_{u\sim\mathbb{Q}(S,\mathbb{P}_S^\theta)}\|u-u^\dagger\|_{\mathcal{U}^{1-\tilde{s}}}^2 \\
=& \frac{1}{2}\tilde{C}^2 M_2^2 \Bigg(\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathbb{E}_{u\sim\mathbb{Q}(S,\mathbb{P}_S^\theta)}\|u-u_p\|_{\mathcal{U}^{1-\tilde{s}}}^2 \\
&\qquad + \mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathbb{E}_{u\sim\mathbb{Q}(S,\mathbb{P}_S^\theta)}\|u_p-u^\dagger\|_{\mathcal{U}^{1-\tilde{s}}}^2\Bigg) \\
=& \frac{1}{2}\tilde{C}^2 M_2^2 \left(\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathrm{Tr}(\mathcal{C}_p) + \mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathbb{E}_{u\sim\mathbb{Q}(S,\mathbb{P}_S^\theta)}\|u_p-u^\dagger\|_{\mathcal{U}^{1-\tilde{s}}}^2\right).
\end{aligned} \tag{87}$$

For the posterior covariance operator $\mathcal{C}_p$, we should notice that it can be written equivalently as $\mathcal{C}_p = \left(\frac{\beta}{m}(\mathcal{L}_x\mathcal{G})^*\Gamma_m^{-1}\mathcal{L}_x\mathcal{G} + \mathcal{C}_0^{-1}\right)^{-1}$, which indicates $\mathrm{Tr}(\mathcal{C}_p) = \frac{m}{\beta}\mathrm{Tr}(\tilde{\mathcal{C}}_p)$. To obtain our final estimate, we need to analyze the term $\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\mathbb{E}_{u\sim\mathbb{Q}(S,\mathbb{P}_S^\theta)}\|u_p-u^\dagger\|_{\mathcal{U}^{1-\tilde{s}}}^2$ in detail. Relying on the explicit form of u_p, we have

$$\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\|u_p-u^\dagger\|_{\mathcal{U}^{1-\tilde{s}}}^2 \leq \mathrm{I}_{11}+\mathrm{I}_{12}+\mathrm{I}_{13}, \tag{88}$$

where

$$\begin{aligned}
\mathrm{I}_{11} =& 3\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\|f(S;\theta)-u^\dagger\|_{\mathcal{U}^{1-\tilde{s}}}^2, \\
\mathrm{I}_{12} =& 3\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\|\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*\left(\frac{m}{\beta}\Gamma_m + \mathcal{L}_{\boldsymbol{x}}\mathcal{G}\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*\right)^{-1}\boldsymbol{\eta}\|_{\mathcal{U}^{1-\tilde{s}}}^2, \\
\mathrm{I}_{13} =& 3\mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{E}_{S\sim\mathbb{D}^m}\|\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*\left(\frac{m}{\beta}\Gamma_m + \mathcal{L}_{\boldsymbol{x}}\mathcal{G}\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*\right)^{-1}(\mathcal{L}_{\boldsymbol{x}}\mathcal{G}(u^\dagger - f(S;\theta)))\|_{\mathcal{U}^{1-\tilde{s}}}^2.
\end{aligned}$$

Relying on the assumptions of f and $\mathscr{E}$, we have the following estimate for term I_{11} and I_{13}

$$\begin{aligned}
\mathrm{I}_{11} \leq& 12R_u^2, \\
\mathrm{I}_{13} \leq& \frac{12\beta^2}{m^2}R_u^2\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}_1^m}\|\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*\left(\Gamma_m + \frac{\beta}{m}\mathcal{L}_{\boldsymbol{x}}\mathcal{G}\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*\right)^{-1}\mathcal{L}_{\boldsymbol{x}}\mathcal{G}\|_{\mathcal{B}(\mathcal{U}^{1-\tilde{s}})}^2.
\end{aligned} \tag{89}$$

Remembering our assumptions on $\boldsymbol{\eta}$, we obtain

$$\mathrm{I}_{12} \leq \frac{3\beta^2}{m^2}\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}_1^m}\mathrm{Tr}(K_{\boldsymbol{x}}\Gamma_m K_{\boldsymbol{x}}^*), \tag{90}$$

where $K_{\boldsymbol{x}} = \mathcal{C}_0^{\frac{1+\tilde{s}}{2}}(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*(\Gamma_m + \frac{\beta}{m}\mathcal{L}_{\boldsymbol{x}}\mathcal{G}\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*)^{-1}$. Combining estimates from (**??**) to (**??**), we obtain the desired result. □

Proof of Lemma 2 (Estimate Sub-Gaussian Parameters for the General Nonlinear Inverse Problem)

Proof. **Calculate** s_{I}^2. The calculation of s_{I}^2 is similar to the first part proof of Lemma 1. The only difference is the estimate of I_1 defined in the proof of Lemma 1. Hence we will only provide the different parts in the following. The term I_1 is defined as follow:

$$I_1 = \frac{1}{2}\mathbb{E}_{x\sim\mathbb{D}_1}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x(\mathcal{G}(u)-\mathcal{G}(u^\dagger))\|_{\mathcal{H}}^2 - \frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_{x_i}(\mathcal{G}(u_i)-\mathcal{G}(u^\dagger))\|_{\mathcal{H}}^2.$$

Obviously, we have $\mathbb{E}_{x_i\sim\mathbb{D}} I_1 = 0$. Let us denote $\tilde{C} := \|\mathcal{C}_1^{\frac{s}{2}}\|_{\mathcal{B}(\mathcal{H})}$. According to the truncated Gaussian prior measure assumption and the compactly supported condition (**??**) of $\mathcal{E}$ in the main text, we find that

$$\begin{aligned}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x(\mathcal{G}(u_i)-\mathcal{G}(u_i^\dagger))\|_{\mathcal{H}}^2 &\le \tilde{C}^2\|\mathcal{C}_1^{-\frac{\tilde{s}}{2}}\Gamma^{-\frac{1}{2}}\mathcal{L}_x(\mathcal{G}(u_i)-\mathcal{G}(u_i^\dagger))\|_{\mathcal{H}}^2\\ &\le \tilde{C}^2 M_2(\|u_i\|_{\mathcal{U}^{1-\tilde{s}}}, \|u_i^\dagger\|_{\mathcal{U}^{1-\tilde{s}}})^2\|u_i-u_i^\dagger\|_{\mathcal{U}^{1-\tilde{s}}}^2\\ &\le 4\tilde{C}^2 M_2(R_u,R_u)^2 R_u^2,\end{aligned}$$

which yields

$$-2\tilde{C}^2 M_2(R_u,R_u)^2R_u^2 \le \mathrm{I}_1 \le 2\tilde{C}^2 M_2(R_u,R_u)^2R_u^2.$$

Now, relying on the Hoeffding's lemma, we find that

$$\mathbb{E}_{\mathbb{D}_i\sim\mathcal{T}}\mathbb{E}_{x_i\sim\mathbb{D}_{i1}}e^{\tilde{\gamma}\mathrm{I}_1} \le \exp\left(\frac{\tilde{\gamma}^2}{2}\tilde{C}^2 M_2(R_u,R_u)^2R_u^2\right). \tag{91}$$

Combining estimates (34) with the estimates (**??**) and (**??**) given in the proof of Lemma 1, we obtain the required estimate.

Calculate s_{II}^2. Similar to the derivation of (**??**), we have

$$\begin{aligned}\mathcal{L}(\mathbb{Q}(S_i,\mathbb{P}_{R_u}^{S_i,\theta}),\mathbb{D}_i) =& \mathbb{E}_{u\sim\mathbb{Q}(S_i,\mathbb{P}_{R_u}^{S_i,\theta})}\mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x(\mathcal{G}(u)-\mathcal{G}(u_i^\dagger))\|_{\mathcal{H}}^2\right]\\ &- \mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u_i^\dagger)\|_{\mathcal{H}}^2\right],\end{aligned}$$

where $\mathbb{P}_{R_u}^{S,\theta}$ is the truncated Gaussian prior measure defined in Remark 8. Relying on the above formulation, we obtain

$$V_i^2 \le \mathbb{E}_{\mathbb{D}_i\sim\mathcal{T}}\mathbb{E}_{S_i\sim\mathbb{D}_i^m}\mathbb{E}_{\theta\sim\mathscr{P}}\exp\left(\tilde{\lambda}\mathrm{I}_1+\tilde{\lambda}\mathrm{I}_2\right), \tag{92}$$

where

$$\begin{aligned}\mathrm{I}_1 =& \mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{D}_{S\sim\mathbb{D}^m}\mathbb{E}_{u\sim\mathbb{Q}(S,\mathbb{P}_{R_u}^{S,\theta})}\mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x(\mathcal{G}(u)-\mathcal{G}(u^\dagger))\|_{\mathcal{H}}^2\right]\\ &- \mathbb{E}_{u\sim\mathbb{Q}(S_i,\mathbb{P}_{R_u}^{S_i,\theta})}\mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x(\mathcal{G}(u)-\mathcal{G}(u_i^\dagger))\|_{\mathcal{H}}^2\right],\\ \mathrm{I}_2 =& \mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u_i^\dagger)\|_{\mathcal{H}}^2\right] - \mathbb{E}_{\mathbb{D}\sim\mathcal{T}}\mathbb{D}_{S\sim\mathbb{D}^m}\mathbb{E}_{x\sim\mathbb{D}_1}\left[\frac{1}{2}\|\Gamma^{-\frac{1}{2}}\mathcal{L}_x\mathcal{G}(u^\dagger)\|_{\mathcal{H}}^2\right].\end{aligned}$$

At this point, we should notice that the posterior measure $\mathbb{Q}(S, \mathbb{P}_{R_u}^{S,\theta})$ obtained from the Bayes' formula with the truncated Gaussian measure has compact support which means $\operatorname{supp}\mathbb{Q}(S, \mathbb{P}_{R_u}^{S,\theta}) \subset B_{\tilde{s}}(R_u)$. Then we have

$$\begin{aligned} -2\tilde{C}^2 M_1(R_u)^2 \leq \mathrm{I}_1 \leq 2\tilde{C}^2 M_1(R_u)^2, \\ -\frac{1}{2}\tilde{C}^2 M_1(R_u)^2 \leq \mathrm{I}_2 \leq \frac{1}{2}\tilde{C}^2 M_1(R_u)^2. \end{aligned} \tag{93}$$

Combining the above estimates (**??**) and estimate (**??**), we could apply the Hoeffding's lemma to obtain

$$V_i^2 \leq \exp\left(\frac{\tilde{\lambda}^2 25\tilde{C}^4 M_1(R_u)^4}{2}\right),$$

which is just the desired result. □

Proof of Theorem 3 (Bayes' Formula for the Hyper-Posterior)

Proof. Here we only provide details for the case of the unbounded loss function since the bounded loss function case is actually easier to be proved. Recall the proofs of Theorem 15 in Dashti & Stuart (2017), we know that the key point is to prove $0 < Z_p < \infty$. Through a trivial calculation, we have

$$Z_p = \int_\Theta \prod_{i=1}^n \left[\int_{\mathcal{U}} \exp\Big(-\sum_{j=1}^m \Phi(u; z_{ij}) \Big) \mathbb{P}_{S_i}^\theta(du) \right]^{\frac{1}{m+1}} \mathscr{P}(d\theta). \tag{94}$$

In the following, we assume the potential function is defined by

$$\Phi(u; z_{ij}) = \frac{1}{2}\|\Gamma^{-\frac{1}{2}} \mathcal{L}_{x_{ij}} \mathcal{G}(u)\|_{\mathcal{H}}^2 - \langle \Gamma^{-\frac{1}{2}} y_{ij}, \Gamma^{-\frac{1}{2}} \mathcal{L}_{x_{ij}} \mathcal{G}(u) \rangle_{\mathcal{H}}.$$

For the potential function defined as in Remark 7, similar techniques can be used to derive the desired results. As in the proof of Theorem 2, we find that for a small enough constant δ_1 such that

$$\sum_{j=1}^m \Phi(u; z_{ij}) \geq -\sum_{j=1}^m \frac{1}{4\delta_1}\|y_{ij}\|_{\mathcal{H}^{\alpha-s}}^2 - \delta_1 m M_1(\|u\|_{\mathcal{U}^{1-\tilde{s}}})^2. \tag{95}$$

Obviously, we have

$$\frac{d\mathbb{P}_{S_i}^\theta}{d\mathbb{P}_0}(u) = \exp\Big(-\tilde{\Phi}(u; S_i) \Big),$$

where $\tilde{\Phi}(u; S_i) = \langle \mathcal{C}_0^{-\frac{1}{2}} f(S_i; \theta), \mathcal{C}_0^{-\frac{1}{2}}(u - f(S_i; \theta)) \rangle_{\mathcal{U}} + \frac{1}{2}\|\mathcal{C}_0^{-\frac{1}{2}} f(S_i; \theta)\|_{\mathcal{U}}^2$. Exactly the same as for the proof of Theorem 2, we derive

$$-\tilde{\Phi}(u; S_i) \leq \exp\Big(\delta_2 \|u\|_{\mathcal{U}^{1-\tilde{s}}}^2\Big) + \left(\frac{1}{4\delta_2} - \frac{3}{2}\right) R_u^2. \tag{96}$$

Combining the estimates (36) and (37), we find

$$\int_{\mathcal{U}} \exp\bigg(-\sum_{j=1}^{m} \Phi(u; z_{ij})\bigg)\mathbb{P}^{\theta}_{S_i}(du) \leq \exp\bigg(\sum_{j=1}^{m} \frac{\|y_{ij}\|^2_{\mathcal{H}^{\alpha-s}}}{4\delta_1} + \bigg(\frac{1}{4\delta_2} - \frac{3}{2}\bigg) R_u^2\bigg) F, \tag{97}$$

where

$$F = \int_{\mathcal{U}} \exp\bigg(\delta_1 m M_1(\|u\|_{\mathcal{U}^{1-\tilde{s}}})^2 + \delta_2 \|u\|^2_{\mathcal{U}^{1-\tilde{s}}}\bigg)\mathbb{P}_0(du) < +\infty.$$

Plugging estimate (**??**) into the formula of Z_p, we derive $Z_p < \infty$. Using the assumptions on the forward operator and the function $f(S;\theta)$, we have

$$\begin{aligned}
\sum_{j=1}^{m} \Phi(u; z_{ij}) \leq & \frac{1}{2}\|\mathcal{C}_0^{\frac{s}{2}}\|^2_{\mathcal{B}(\mathcal{U})} M_1(\|u\|_{\mathcal{U}^{1-\tilde{s}}})^2 + M_1(\|u\|_{\mathcal{U}^{1-\tilde{s}}}) \sum_{j=1}^{m} \|y_{ij}\|_{\mathcal{H}^{\alpha-s}} \\
\tilde{\Phi}(u; S_i) \leq & \frac{3}{2} R_u^2 + R_u \|u\|_{\mathcal{U}^{1-\tilde{s}}}.
\end{aligned} \tag{98}$$

Using estimates in (**??**), we derive

$$\int_{\mathcal{U}} \exp\bigg(-\sum_{j=1}^{m} \Phi(u; z_{ij})\bigg)\mathbb{P}^{\theta}_{S_i}(du) > 0 \tag{99}$$

independent of the parameter θ, which yields $Z_p > 0$. □

Proof of Theorem 4 (The Maximum a Posterior Estimate)

Proof. Let us denotes

$$\begin{aligned}
\Phi^h(\theta) :=& \frac{-1}{m+1} \sum_{i=1}^{n} \ln Z_m(S_i, \mathbb{P}^{\theta}_{S_i}) \\
=& \frac{-1}{m+1} \sum_{i=1}^{n} \ln \int_{\mathcal{U}} \exp\bigg(-\sum_{j=1}^{m} \Phi(u; z_{ij}) - \tilde{\Phi}(u; S_i, \theta)\bigg)\mathbb{P}_0(du),
\end{aligned} \tag{100}$$

where

$$\begin{aligned}
\Phi(u; z_{ij}) :=& \frac{1}{2}\|\Gamma^{-1/2}\mathcal{L}_{x_{ij}}\mathcal{G}(u)\|^2_{\mathcal{H}} - \langle \Gamma^{-1/2} y_{ij}, \Gamma^{-1/2}\mathcal{L}_{x_{ij}}\mathcal{G}(u)\rangle_{\mathcal{H}}, \\
\tilde{\Phi}(u; S_i, \theta) :=& -\ln \frac{d\mathbb{P}^{\theta}_{S_i}}{d\mathbb{P}_0}(u) \\
=& \frac{1}{2}\|\mathcal{C}_0^{-1/2} f(S_i;\theta)\|^2_{\mathcal{U}} - \langle \mathcal{C}_0^{-1/2} f(S_i;\theta), \mathcal{C}_0^{-1/2}(u - f(S_i;\theta))\rangle_{\mathcal{U}}.
\end{aligned}$$

Similar to the proof of Theorem 3 of the main text, we will know that Φ^h is locally bounded from above and below with respect to the parameter θ. So the main point is to verify the local Lipschitz continuity of Φ^h with respect to θ. Obviously, the Fréchet derivative of Φ^h is

$$D_\theta \Phi^h = \sum_{i=1}^{n} \frac{\int_{\mathcal{U}} \exp\Big(-\sum_{j=1}^{m} \Phi(u; z_{ij}) - \tilde{\Phi}(u; S_i, \theta)\Big) D_\theta \tilde{\Phi}(u; S_i, \theta)\mathbb{P}_0(du)}{(m+1)\int_{\mathcal{U}} \exp\Big(-\sum_{j=1}^{m} \Phi(u; z_{ij})\Big)\mathbb{P}^{\theta}_{S_i}(du)}.$$

Recall the following assumption

$$\|D_\theta \tilde{\Phi}(u; S_i, \theta)\|_{\Theta^*} \leq M_5(r)\|u\|_{\mathcal{U}^{1-\tilde{s}}},$$

where $\theta \in B_\Theta(\delta) := \{\theta \in \Theta : \|\theta\|_\Theta < r\}$ and $M_5(r)$ is a constant depends on r. Then we have

$$\begin{aligned}\|D_\theta \Phi^h\|_{\Theta^*} \leq & \sum_{i=1}^n \frac{\int_{\mathcal{U}} \exp\left(-\sum_{j=1}^m \Phi(u; z_{ij}) - \tilde{\Phi}(u; S_i, \theta)\right) \|D_\theta \tilde{\Phi}(u; S_i, \theta)\|_{\Theta^*} \mathbb{P}_0(du)}{(m+1)\int_{\mathcal{U}} \exp\left(-\sum_{j=1}^m \Phi(u; z_{ij})\right) \mathbb{P}^\theta_{S_i}(du)} \\ \leq & \sum_{i=1}^n \frac{M_5(r)\int_{\mathcal{U}} \exp\left(-\sum_{j=1}^m \Phi(u; z_{ij}) - \tilde{\Phi}(u; S_i, \theta)\right) \|u\|_{\mathcal{U}^{1-\tilde{s}}} \mathbb{P}_0(du)}{(m+1)\int_{\mathcal{U}} \exp\left(-\sum_{j=1}^m \Phi(u; z_{ij})\right) \mathbb{P}^\theta_{S_i}(du)}. \end{aligned} \tag{101}$$

For the numerator term on the right-hand side of (39), it can be bounded from above by employing similar techniques as for deriving estimate (**??**). The only difference is the extra term $\|u\|_{\mathcal{U}^{1-\tilde{s}}}$ that can be bounded by $\|u\|_{\mathcal{U}^{1-\tilde{s}}} \leq C\exp(\epsilon\|u\|_{\mathcal{U}^{1-\tilde{s}}})$ with arbitrarily small constant $\epsilon > 0$. Hence, the upper bound is obtained by taking $\delta_2 + \epsilon$ (δ_2 is the same as in (**??**)) small enough. For the denominator term on the right-hand side of (39), it is exactly the same as that of (**??**) with a positive lower bound independent of θ. Now we can conclude that

$$\|D_\theta \Phi^h\|_{\Theta^*} \leq C(r) < +\infty,$$

where $C(r)$ is some constant depends on r. This obviously indicates Φ^h is locally Lipschitz continuous with respect to the parameter θ. Now, following the illustrations given in Subsection 3.1 of Trillos & Sanz-Alonso (2020), we will obtain the desired results. □

B.2 Proof Details of Appendix A.6

Verify Condition (??) for the Backward Diffusion Problem

Proof. Since the operator $\mathcal{C}_0$ has the following eigen-system decomposition:

$$\mathcal{C}_0 e_k = \lambda_k e_k, \quad \text{for all } k = 1, 2, \ldots, \tag{102}$$

we easily find that

$$Ae_k = \sqrt{\frac{\lambda}{\lambda_k}} e_k, \quad \mathcal{C}_1 e_k = \frac{\lambda_k}{\lambda} e_k, \quad \text{for all} k = 1, 2, \ldots. \tag{103}$$

Employing (40) and (41), we obtain

$$\begin{aligned}\|\mathcal{C}_1^{-\frac{s}{2}}\Gamma^{-\frac{1}{2}} e^{-AT} u\|^2_{\mathcal{H}} =& \tau^{-2}\left\|\sum_{k=1}^\infty u_k \left(\frac{\lambda}{\lambda_k}\right)^{\frac{s}{2}} e^{-T\sqrt{\frac{\lambda}{\lambda_k}}} e_k\right\|^2_{\mathcal{H}} \\ =& \tau^{-2}\sum_{k=1}^\infty u_k^2 \frac{\lambda^s}{\lambda_k^s} e^{-2T\sqrt{\frac{\lambda}{\lambda_k}}} \\ =& \tau^{-2}\sum_{k=1}^\infty \lambda_k^{-(1-\tilde{s})} u_k^2 \lambda_k^{1-s} \frac{\lambda^s}{\lambda_k^s} e^{-2T\sqrt{\frac{\lambda}{\lambda_k}}} \\ \leq & \tau^{-2}\left(\sup_k \lambda^s \lambda_k^{1-2s} e^{-2T\sqrt{\frac{\lambda}{\lambda_k}}}\right)\|u\|^2_{\mathcal{U}^{1-\tilde{s}}},\end{aligned}$$

which is the required estimate. □

Calculate s_I and s_{II} in (??) for the Backward Diffusion Problem

Proof. For every $u \in \mathcal{H}$, denote $u_k = \langle u, e_k \rangle_{\mathcal{H}}$, we have

$$\|\mathcal{C}_0^{\frac{s}{2}} u\|_{\mathcal{H}}^2 = \sum_{k=1}^{\infty} \lambda_k^s u_k^2 \leq \lambda_1^s \|u\|_{\mathcal{H}}^2.$$

Taking $u = u_1 e_1$, we obtain $\|\mathcal{C}_0^{\frac{s}{2}} u\|_{\mathcal{H}}^2 \geq \lambda_1^s \|u\|_{\mathcal{H}}^2$, which yields $\tilde{C} = \lambda^{-s/2} \|\mathcal{C}_0^{\frac{s}{2}}\|_{\mathcal{B}(\mathcal{H})} = \left(\frac{\lambda_1}{\lambda}\right)^{s/2}$. Plugging $\tilde{C}$, M_1, and M_2 into the equation of s_{I}^2, we have

$$\begin{aligned} s_{\mathrm{I}}^2 =& \frac{\lambda_1^s}{\lambda^s \tau^2} \sup_k \left[\lambda^s \lambda_k^{1-2s} e^{-2T\sqrt{\frac{\lambda}{\lambda_k}}} \right] \left(8R_u^2 + \mathrm{Tr}(\mathcal{C}_1^s) + \frac{\lambda_1^s \sup_k \left[\lambda^s \lambda_k^{1-2s} e^{-2T\sqrt{\frac{\lambda}{\lambda_k}}} \right]}{4\lambda^s \tau^2} \right) \\ & - \frac{1}{2} \ln \det \left(\mathrm{Id} - 2 \frac{\lambda_1^s}{\lambda^2 \tau^2} \sup_k \left[\lambda^s \lambda_k^{1-2s} e^{-2T\sqrt{\frac{\lambda}{\lambda_k}}} \right] \mathcal{C}_0 \right) \end{aligned}$$

Concerned with s_{II}^2, we firstly give the following two estimates:

$$\begin{aligned} \frac{m}{\beta} \mathrm{Tr}(\tilde{\mathcal{C}}_p) =& \frac{m}{\beta} \sum_{k=1}^{\infty} \langle \tilde{\mathcal{C}}_p e_k, e_k \rangle_{\mathcal{H}} = \frac{m}{\beta} \sum_{k=1}^{\infty} \frac{1}{\tau^{-2} m e^{-2\lambda_k^{-1/2} T} + \frac{m}{\beta} \lambda_k^{-1}} \\ =& \sum_{k=1}^{\infty} \frac{1}{\tau^{-2} \beta e^{-2\lambda_k^{-1/2} T} + \lambda_k^{-1}} \leq \sum_{k=1}^{\infty} \lambda_k = \mathrm{Tr}(\mathcal{C}_0), \end{aligned} \tag{104}$$

and

$$\begin{aligned} \frac{3\beta^2}{m^2} \mathbb{E}_{\boldsymbol{x} \sim \mathbb{D}_1^m} \mathrm{Tr}(K_{\boldsymbol{x}} \Gamma_m K_{\boldsymbol{x}}^*) =& \tau^2 \frac{3\beta^2}{m^2} \sum_{k=1}^{\infty} \langle K_{\boldsymbol{x}} K_{\boldsymbol{x}}^* e_k, e_k \rangle_{\mathcal{H}} \\ =& \tau^2 \frac{3\beta^2}{m^2} \sum_{k=1}^{\infty} \lambda_k^{1+s} e^{-2\sqrt{\frac{\lambda}{\lambda_k}} T} \left(\tau^2 + \frac{\beta}{m} e^{-2\sqrt{\frac{\lambda}{\lambda_k}} T} \lambda_k \right)^{-2} \\ \leq& \frac{3\beta}{m} \sum_{k=1}^{\infty} \lambda_k^s \\ =& \frac{3\beta}{m} \mathrm{Tr}\left(\mathcal{C}_0^s\right). \end{aligned} \tag{105}$$

From the following estimate

$$\begin{aligned} & \|\mathcal{C}_0 (\mathcal{L}_{\boldsymbol{x}} \mathcal{G})^* (\Gamma_m + \frac{\beta}{m} \mathcal{L}_{\boldsymbol{x}} \mathcal{G} \mathcal{C}_0 (\mathcal{L}_{\boldsymbol{x}} \mathcal{G})^*)^{-1} \mathcal{L}_{\boldsymbol{x}} \mathcal{G} u\|_{\mathcal{H}^{1-s}}^2 \\ =& \sum_{k=1}^{\infty} \left[\lambda_k e^{-2\sqrt{\frac{\lambda}{\lambda_k}} T} \left(\tau^2 + \frac{\beta}{m} e^{-2\sqrt{\frac{\lambda}{\lambda_k}} T} \lambda_k \right)^{-1} \right]^2 \lambda_k^{s-1} u_k^2 \\ \leq& \frac{m^2}{\beta^2} \|u\|_{\mathcal{H}^{1-s}}^2, \end{aligned}$$

we derive

$$\frac{12\beta^2}{m^2}R_u^2\mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}_{\Gamma}^m}\|\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*(\Gamma_m+\frac{\beta}{m}\mathcal{L}_{\boldsymbol{x}}\mathcal{G}\mathcal{C}_0(\mathcal{L}_{\boldsymbol{x}}\mathcal{G})^*)^{-1}\mathcal{L}_{\boldsymbol{x}}\mathcal{G}\|^2_{\mathcal{B}(\mathcal{U}^{1-\tilde{s}})}\leq 12R_u^2. \tag{106}$$

Combining estimates (42), (43), and (44), we obtain the desired formula of s_{II}^2, which completes the proof. □

References

Agapiou, S., Larsson, S. & Stuart, A. M. (2013), 'Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems', *Stoch. Proc. Appl.* **123**(10), 3828–3860.

Alquier, P. (2024), 'User-friendly introduction to PAC-Bayes bounds', *Foundations and Trends in Machine Learning* **17**(2), 174–303.

Alquier, P., Ridgway, J. & Chopin, N. (2016), 'On the properties of variational approximations of Gibbs posteriors', *J. Mach. Learn. Res.* **17**, 1–41.

Amit, R. & Meir, R. (2018), Meta-learning by adjusting priors based on extended PAC-Bayes theory, *in* 'Proceedings of the International Conference on Machine Learning', Vol. 80, pp. 205–214.

Anandkumar, A., Azizzadenesheli, K., Bhattacharya, K., Kovachki, N., Li, Z., Liu, B. & Stuart, A. M. (2020), Neural operator: Graph kernel network for partial differential equations, *in* 'ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations'.

Bao, G., Li, P., Lin, J. & Triki, F. (2015), 'Inverse scattering problems with multi-frequencies', *Inverse Probl.* **31**(9), 093001.

Bhattacharya, K., Hosseini, B., Kovachki, N. B. & Stuart, A. M. (2021), 'Model reduction and neural networks for parametric PDEs', *SMAI J. Comput. Math.* **7**, 121–157.

Bottou, L., Curtis, F. E. & Nocedal, J. (2018), 'Optimization methods for large-scale machine learning', *SIAM Rev.* **60**(2), 223–311.

Bui-Thanh, T., Ghattas, O., Martin, J. & Stadler, G. (2013), 'A computational framework for infinite-dimensional Bayesian inverse problems part I', *SIAM J. Sci. Comput.* **35**(6), A2494–A2523.

Bui-Thanh, T. & Nguyen, Q. P. (2016), 'FEM-based discretization-invariant MCMC methods for PDE-constrained Bayesian inverse problems', *Inverse Probl. Imag.* **10**(4), 943–975.

Cotter, S. L., Roberts, G. O., Stuart, A. M. & White, D. (2013), 'MCMC methods for functions: modifying old algorithms to make them faster', *Stat. Sci.* **28**(3), 424–446.

Dashti, M. & Stuart, A. M. (2017), 'The Bayesian approach to inverse problems', *Handbook of Uncertainty Quantification* pp. 311–428.

Engl, H. W., Hanke, M. & Neubauer, A. (1996), *Regularization of Inverse Problems*, Springer, Netherlands.

Germain, P., Bach, F., Lacoste, A. & Lacoste-Julien, S. (2016), PAC-Bayesian theory meets Bayesian inference, *in* 'Advances in Neural Information Processing Systems', Vol. 29, p. 1884–1892.

Guedj, B. (2019), 'A primer on PAC-Bayesian learning', arXiv:1901.05353.

Ito, K. & Jin, B. (2015), *Inverse Problems: Tikhonov Theory and Algorithms*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ.

Jia, J., Li, P. & Meng, D. (2022), 'Stein variational gradient descent on infinite-dimensional space and applications to statistical inverse problems', *SIAM J. Numer. Anal.* **60**(4), 2225–2252.

Jia, J., Peng, J. & Gao, J. (2016), 'Bayesian approach to inverse problems for functions with a variable-index Besov prior', *Inverse Probl.* **32**(8), 085006.

Jia, J., Peng, J., Gao, J. & Li, Y. (2018), 'Backward problem for a time-space fractional diffusion equation', *Inverse Probl. Imag.* **12**(3), 773–799.

Jia, J., Peng, J. & Yang, J. (2017), 'Harnack's inequality for a space-time fractional diffusion equation and applications to an inverse source problem', *J. Differ. Equations* **262**(8), 4415–4450.

Jia, J., Wu, B., Peng, J. & Gao, J. (2019), 'Recursive linearization method for inverse medium scattering problems with complex mixture gaussian error learning', *Inverse Probl.* **35**(7), 075003.

Jia, J., Yue, S., Peng, J. & Gao, J. (2018), 'Infinite-dimensional Bayesian approach for inverse scattering problems of a fractional Helmholtz equation', *J. Funct. Anal.* **275**(9), 2299–2332.

Jia, J., Zhao, Q., Xu, Z., Meng, D. & Leung, Y. (2021), 'Variational Bayes' method for functions with applications to some inverse problems', *SIAM J. Sci. Comput.* **43**(1), A355–A383.

Jin, B. & Rundell, W. (2015), 'A tutorial on inverse problems for anomalous diffusion processes', *Inverse Probl.* **31**(3), 035003.

Jin, B. & Zou, J. (2010), 'Hierarchical Bayesian inference for ill-posed problems via variational method', *J. Comput. Phys.* **229**(19), 7317–7343.

Kovachki, N., Lanthaler, S. & Mishra, S. (2021), 'On universal approximation and error bounds for fourier neural operators', *J. Mach. Learn. Res.* **22**, 1–76.

Li, J., Yammamoto, M. & Zou, J. (2009), 'Conditional stability and numerical reconstruction of initial temperature', *Commun. Pur. Appl. Math.* **8**(1), 361–382.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A. M. & Anandkumar, A. (2020), Fourier neural operator for parametric partial differential equations, *in* 'International Conference on Learning Representations'.

Liu, Q. & Wang, D. (2016), Stein variational gradient descent: A general purpose Bayesian inference algorithm, *in* 'Advances in Neural Information Processing Systems', Vol. 29.

Logg, A., Mardal, K. A. & Wells, G. N. (2012), *Automated Solution of Differential Equations by the Finite Element Method*, Springer, Heidelberg.

McAllester, D. A. (1998), Some pac-bayesian theorems, *in* 'Proceedings of the 11th Annual Conference on Computational Learning Theory', pp. 230–234.

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2018), *Foundations of Machine Learning*, The MIT Press, London, England.

Monard, F., Nickl, R. & Paternain, G. P. (2021), 'Consistent inversion of noisy non-Abelian X-ray transforms', *Commun. Pur. Appl. Math.* **74**(5), 1045–1099.

Nelsen, N. H. & Stuart, A. M. (2021), 'The random feature model for input-output maps between Banach spaces', *SIAM J. Sci. Comput.* **43**(5), A3212–A3243.

Nickl, R. (2020), 'Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation', *J. Eur. Math. Soc.* **22**(8), 2697–2750.

Nickl, R. (2023), *Bayesian Non-Linear Statistical Inverse Problems*, EMS Press, Berlin.

Nickl, R. & Söhl, J. (2017), 'Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions', *Ann. Statist.* **45**(4), 1664 – 1693.

Pazy, A. (1983), *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York.

Pentina, A. & Lampert, C. (2014), A PAC-Bayesian bound for life-long learning, *in* 'Proceedings of the International Conference on Machine Learning', Vol. 32, pp. 991–999.

Pinski, F. J., Simpson, G., Stuart, A. M. & Weber, H. (2015*a*), 'Algorithms for Kullback–Leibler approximation of probability measures in infinite dimensions', *SIAM J. Sci. Comput.* **37**(6), A2733–A2757.

Pinski, F. J., Simpson, G., Stuart, A. M. & Weber, H. (2015*b*), 'Kullback-Leibler approximation for probability measures on infinite dimensional space', *SIAM J. Math. Anal.* **47**(6), 4091–4122.

Prato, G. D. (2006), *An Introduction to Infinite-Dimensional Analysis*, Springer-Verlag, Berlin.

Prato, G. D. & Zabczyk, J. (2014), *Stochastic Equations in Infinite Dimensions*, second edn, Cambridge University Press, Cambridge.

Richter, G. R. (1981), 'An inverse problem for the steady state diffusion equation', *SIAM J. Appl. Math.* **41**(2), 210–221.

Rothfuss, J., Fortuin, V., Josifoski, M. & Krause, A. (2021), PACOH: Bayes-optimal meta-learning with PAC-guarantees, *in* 'Proceedings of the International Conference on Machine Learning', Vol. 139, pp. 9116–9126.

Stuart, A. M. (2010), 'Inverse problems: A Bayesian perspective', *Acta Numer.* **19**, 451–559.

Trillos, N. G. & Sanz-Alonso, D. (2020), 'The Bayesian update: Variational formulations and gradient flow', *Bayesian Anal.* **15**(1), 29–56.

Vollmer, S. J. (2013), 'Posterior consistency for Bayesian inverse problems through stability and regression results', *Inverse Probl.* **29**(12), 125011.