

In appendix A we provide additional details and further results of experiments. In appendix B, we list the assumptions we used, and prove the non-asymptotic version of Theorem 4.2 and 4.6. In appendix C, we give the details of Sec. 3, including deriving algorithms presented in Sec. 3.2, examples in Sec. 3.4 and a general formula for curl-free kernels. Appendix D includes some technical results used in proofs. Finally, We present samples drawn from trained WAEs in appendix E.

A. Experiment Details and Additional Results

In experiments, we use the IMQ kernel $k(\mathbf{x}, \mathbf{y}) := (1 + \|\mathbf{x} - \mathbf{y}\|_2^2 / \sigma^2)^{-1/2}$ and its curl-free version in corresponding kernel estimators. We use the median of the pairwise Euclidean distances between samples as the kernel bandwidth. The parameter ν of the ν -method is set to 1. The maximum iteration number of KEF-CG is 40 and the convergence tolerance of it is 10^{-4} .

A.1. Grid Distributions

We use αM eigenvalues in SSGE with α searched in $\{0.99, 0.97, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4\}$. We search the number of iteration steps of the ν -method in $\{20, 30, 40, 50, 60, 70, 80, 90, 100\}$. We search the regularization coefficient λ of Stein, NKEF, KEF-CG in $\{10^{-k} : k = 0, 1, \dots, 8\}$. The experiments are repeated 32 times.

A.2. Wasserstein Autoencoders

We use the standard Gaussian distribution $\mathcal{N}(0, I)$ as the prior $p(\mathbf{z})$, and $\mathcal{N}(\mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x}))$ as the approximated posterior $q_\phi(\mathbf{z}|\mathbf{x})$, and Bernoulli($G_\theta(\mathbf{z})$) as the generator $p_\theta(\mathbf{x}|\mathbf{z})$. We use minibatch size 64. Models are optimized by the Adam optimizer with learning rate 10^{-4} . Each configuration is repeated 3 times, and the mean and the standard deviation are reported in Table 3 and Table 4. All models are timed on GeForce GTX TITAN X GPU.

Table 3. Negative log-likelihoods on the MNIST dataset and per epoch time on 128 latent dimension.

LATENT DIM	8	32	64	128	TIME
STEIN	97.15 ± 0.14	92.10 ± 0.07	101.60 ± 0.44	114.41 ± 0.25	4.2s
SSGE	97.24 ± 0.07	92.24 ± 0.17	101.92 ± 0.08	114.57 ± 0.23	9.2s
KEF	97.07 ± 0.03	90.93 ± 0.23	91.58 ± 0.03	92.40 ± 0.34	201.1s
NKEF ₂	97.71 ± 0.24	92.29 ± 0.41	92.82 ± 0.18	94.14 ± 0.69	36.4s
NKEF ₄	97.59 ± 0.15	91.19 ± 0.08	91.80 ± 0.12	92.94 ± 0.58	97.5s
NKEF ₈	97.23 ± 0.06	90.86 ± 0.09	92.39 ± 1.32	92.49 ± 0.41	301.2s
KEF-CG	97.39 ± 0.22	90.77 ± 0.12	92.66 ± 0.67	92.05 ± 0.06	13.7s
ν -METHOD	97.28 ± 0.17	90.94 ± 0.02	91.48 ± 0.09	92.10 ± 0.06	78.1s
SSM	96.98 ± 0.27	89.06 ± 0.01	93.06 ± 0.68	96.92 ± 0.08	6.0s

Table 4. Fréchet Inception Distances on the CelebA dataset and per epoch time on 128 latent dimension.

LATENT DIM	8	32	64	128	TIME
STEIN	73.85 ± 1.39	58.29 ± 0.46	57.54 ± 0.57	76.31 ± 1.33	164.4s
SSGE	72.49 ± 1.09	58.01 ± 0.60	58.39 ± 1.00	76.85 ± 1.12	172.2s
NKEF ₂	75.12 ± 1.55	53.92 ± 0.29	51.16 ± 0.30	55.17 ± 0.43	244.7s
NKEF ₄	73.15 ± 0.77	54.54 ± 1.02	50.76 ± 0.19	53.70 ± 0.10	412.5s
KEF-CG	72.92 ± 0.60	54.32 ± 0.31	50.44 ± 0.20	50.66 ± 0.89	166.2s
ν -METHOD	72.02 ± 1.22	52.86 ± 0.20	50.16 ± 0.23	52.80 ± 0.43	220.9s
SSM	69.72 ± 0.25	49.93 ± 0.74	72.68 ± 1.75	94.07 ± 3.57	163.3s

MNIST We parameterize μ_ϕ , σ_ϕ^2 and $G_\theta(\mathbf{z})$ by fully-connected neural networks with two hidden layers, both of which consist of 256 units activated by ReLU. For SSM, the score is parameterized by a fully-connected neural network with two hidden layers consisting of 256 units activated by tanh. The regularization coefficients of Stein, KEF, NKEF, KEF-CG are searched in $\{10^{-k} : k = 2, 3, \dots, 7\}$ for the best log-likelihood, and the number of iteration steps of the ν -method are searched in $\{50, 70, \dots, 150\}$, and we use αM eigenvalues in SSGE with α searched in $\{0.99, 0.97, 0.95, 0.93, 0.91, 0.89, 0.87\}$. We run 1000 epoches and evaluate the model by AIS (Neal, 2001), where the parameters are the same as in SSGE. Specifically,

we set the step size of HMC to 10^{-6} , and the leapfrog step to 10. We use 5 chains and set the temperature to 10^3 .

CelebA We parameterize μ_ϕ , G_θ by convolutional neural networks similar to Song et al. (2019). σ_ϕ^2 is set to 1. For SSM, we use the same network as in MNIST to parameterize the score. The regularization coefficients of Stein, KEF, NKEF, KEF-CG are searched in $\{10^{-k} : k = 2, 3, \dots, 7\}$ for the best log-likelihood, and the number of iteration steps of the ν -method are searched in $\{20, 30, 40, 50, 60, 70\}$, and we use αM eigenvalues in SSGE with α searched in $\{0.99, 0.97, 0.95, 0.93, 0.91, 0.89, 0.87\}$. We run 100 epoches and evaluate the model using the Fréchet Inception Distance (FID). As KEF and NKEF₈ are slow, we do not compare them in this dataset. Results are reported in Table 4.

B. Error Bounds

In the following, we suppress the dependence of $\mathcal{H}_\mathcal{K}$ on \mathcal{K} for simplicity. We use $\|\cdot\|_{\text{HS}}$ to denote the Hilbert-Schmidt norm of operators. The assumptions required in obtaining an error bound are listed below.

Assumption B.1. \mathcal{X} is a non-empty open subset of \mathbb{R}^d , with piecewise C^1 boundary.

Assumption B.2. p , $\log p$ and each element of \mathcal{K} are continuously differentiable. p and its total derivative $Dp : \mathcal{X} \rightarrow \mathbb{R}^d$ can both be continuously extended to $\bar{\mathcal{X}}$, where $\bar{\mathcal{X}}$ is the closure of \mathcal{X} . Each element of \mathcal{K} and its total derivative can be continuously extended to $\bar{\mathcal{X}} \times \bar{\mathcal{X}}$.

Assumption B.3. For all $i, j \in [d]$, $\mathcal{K}(\mathbf{x}, \mathbf{x})_{ij} p(\mathbf{x}) = 0$ on $\partial\mathcal{X}$, and $\sqrt{|\mathcal{K}(\mathbf{x}, \mathbf{x})_{ij}|} p(\mathbf{x}) = o(\|\mathbf{x}\|_2^{1-d})$ as $\mathbf{x} \rightarrow \infty$, where $\partial\mathcal{X} := \bar{\mathcal{X}} \setminus \mathcal{X}$.

Assumption B.4. Define an $\mathcal{H}_\mathcal{K}$ -valued random variable $\xi_{\mathbf{x}} := \text{div}_{\mathbf{x}} \mathcal{K}_{\mathbf{x}}^T$, let $\xi := \int_{\mathcal{X}} \xi_{\mathbf{x}} d\rho$. There are two constants Σ , K , such that

$$\int_{\mathcal{X}} \left\{ \exp\left(\frac{\|\xi_{\mathbf{x}} - \xi\|_{\mathcal{H}}}{K}\right) - \frac{\|\xi_{\mathbf{x}} - \xi\|_{\mathcal{H}}}{K} - 1 \right\} d\rho \leq \frac{\Sigma^2}{2K^2}.$$

Assumption B.5. There is a constant $\kappa > 0$ such that $\sup_{\mathbf{x} \in \mathcal{X}} \text{tr} \mathcal{K}(\mathbf{x}, \mathbf{x}) \leq \kappa^2$.

Assumptions B.1-B.3 are similar to those in Sriperumbudur et al. (2017). They guarantee the integration by parts is valid, so we can obtain $\mathbb{E}_\rho[\mathcal{K}_{\mathbf{x}} \nabla \log p] = -\mathbb{E}_\rho[\text{div}_{\mathbf{x}} \mathcal{K}_{\mathbf{x}}^T]$. Assumptions B.4 and B.5 come from Bauer et al. (2007), and are used in the concentration inequalities. Note that Assumption B.4 can be replaced by a stronger one that $\|\xi_{\mathbf{x}} - \xi\|_{\mathcal{H}}$ is uniformly bounded on \mathcal{X} .

We follows the idea of Bauer et al. (2007, Theorem 10) to prove Theorem 4.2. The non-asymptotic version is given as follows

Theorem B.1. Assume Assumptions B.1-B.5 hold. Let \bar{r} be the qualification of the regularizer g_λ , and $\hat{s}_{p,\lambda}^g$ be defined as in (8). Suppose there exists $f_0 \in \mathcal{H}_\mathcal{K}$ such that $s_p = L_{\mathcal{K}}^r f_0$, for some $r \in [0, \bar{r}]$. Then for any $0 < \delta < 1$, $M \geq (2\sqrt{2}\kappa^2 \log(4/\delta))^{\frac{2r+2}{r}}$, choosing $\lambda = M^{-\frac{1}{2r+2}}$, the following inequalities hold with probability at least $1 - \delta$

$$\|\hat{s}_{p,\lambda} - s_p\|_{\mathcal{H}} \leq C_1 M^{-\frac{r}{2r+2}} \log \frac{4}{\delta},$$

and for $r \in [0, \bar{r} - 1/2]$, we have

$$\|\hat{s}_{p,\lambda} - s_p\|_\rho \leq C_2 M^{-\frac{2r+1}{4r+4}} \log \frac{4}{\delta},$$

where $C_1 = 2B(K + \Sigma) + 2\sqrt{2}B\kappa^2 \|s_p\|_{\mathcal{H}} + (\gamma_r + \kappa^2 \gamma_{c_r}) \|f_0\|_{\mathcal{H}}$, and $C_2 = 2B(K + \Sigma)\kappa + 2\sqrt{2}B\kappa^3 \|s_p\|_{\mathcal{H}} + ((\gamma_r + \kappa^2 \gamma_{\frac{1}{2}c_r}) + c_{\frac{1}{2}}(\gamma_r + \kappa^2 \gamma_{c_r})) \|f_0\|_{\mathcal{H}}$, and c_r is a constant depending on r . O_p is the Big-O notation in probability.

Proof. We consider the following decomposition

$$\begin{aligned} \|\hat{s}_{p,\lambda} - s_p\|_{\mathcal{H}} &\leq \|g_\lambda(\hat{L}_\mathcal{K})(\hat{\zeta} - \zeta)\|_{\mathcal{H}} + \|g_\lambda(\hat{L}_\mathcal{K})L_\mathcal{K}s_p - s_p\|_{\mathcal{H}} \\ &\leq \|g_\lambda(\hat{L}_\mathcal{K})(\hat{\zeta} - \zeta)\|_{\mathcal{H}} + \|g_\lambda(\hat{L}_\mathcal{K})(L_\mathcal{K} - \hat{L}_\mathcal{K})s_p\|_{\mathcal{H}} + \|r_\lambda(\hat{L}_\mathcal{K})s_p\|_{\mathcal{H}}, \end{aligned}$$

where $r_\lambda(\sigma) := g_\lambda(\sigma)\sigma - 1$. By Definition 4.1, we have $\|g_\lambda(\hat{L}_\mathcal{K})\| \leq B/\lambda$. From Lemma D.3 and D.4, with probability at least $1 - \delta$, we have

$$\|g_\lambda(\hat{L}_\mathcal{K})(\hat{\zeta} - \zeta)\|_{\mathcal{H}} + \|g_\lambda(\hat{L}_\mathcal{K})(L_\mathcal{K} - \hat{L}_\mathcal{K})s_p\|_{\mathcal{H}} \leq \frac{2B(K + \Sigma) + 2\sqrt{2}B\kappa^2 \|s_p\|_{\mathcal{H}}}{\lambda\sqrt{M}} \log \frac{4}{\delta}.$$

By Definition 4.1, $\|r_\lambda(\hat{L}_\mathcal{K})L_\mathcal{K}^r\| \leq \gamma_r\lambda^r$ and $\|r_\lambda(\hat{L}_\mathcal{K})\| \leq \gamma$, then

$$\begin{aligned} \|r_\lambda(\hat{L}_\mathcal{K})s_p\|_{\mathcal{H}} &\leq \|r_\lambda(\hat{L}_\mathcal{K})\hat{L}_\mathcal{K}^r f_0\|_{\mathcal{H}} + \|r_\lambda(\hat{L}_\mathcal{K})(L_\mathcal{K}^r - \hat{L}_\mathcal{K}^r)f_0\|_{\mathcal{H}} \\ &\leq \gamma_r\lambda^r \|f_0\|_{\mathcal{H}} + \gamma\|L_\mathcal{K}^r - \hat{L}_\mathcal{K}^r\| \|f_0\|_{\mathcal{H}}. \end{aligned}$$

When $r \in [0, 1]$, from Bauer et al. (2007, Theorem 1), there exists a constant c_r such that $\|L_\mathcal{K}^r - \hat{L}_\mathcal{K}^r\| \leq c_r\|L_\mathcal{K} - \hat{L}_\mathcal{K}\|^r$. Then by Lemma D.4, and choose $\lambda \geq 2\sqrt{2}\kappa^2 M^{-1/2} \log(4/\delta)$, we have

$$\|L_\mathcal{K}^r - \hat{L}_\mathcal{K}^r\| \leq c_r \left(\frac{2\sqrt{2}\kappa^2 \log \frac{4}{\delta}}{\sqrt{M}} \right)^r \leq c_r\lambda^r.$$

Collecting the above results,

$$\|\hat{s}_{p,\lambda} - s_p\|_{\mathcal{H}} \leq \left(\frac{A_1}{\lambda\sqrt{M}} + A_2\lambda^r \right) \log \frac{4}{\delta},$$

where A_1, A_2 are constants which do not depend on λ and M . Then, we can choose $\lambda = M^{-\frac{1}{2r+2}}$ to obtain the bound. Combining with $\lambda \geq 2\sqrt{2}\kappa^2 M^{-1/2} \log(4/\delta)$, we require $M^{\frac{1}{2r+2}} \geq 2\sqrt{2}\kappa^2 \log(4/\delta)$.

When $r > 1$, from Lemma D.5, there exists a constant c'_r such that $\|L_\mathcal{K}^r - \hat{L}_\mathcal{K}^r\|_{\text{HS}} \leq c'_r\|L_\mathcal{K} - \hat{L}_\mathcal{K}\|_{\text{HS}}$. Then $\|L_\mathcal{K}^r - \hat{L}_\mathcal{K}^r\|_{\text{HS}} \leq 2\sqrt{2}c'_r\kappa^2 M^{-1/2} \log(4/\delta)$, and a similar discussion can be applied to obtain the bound.

Note that $\|\hat{s}_{p,\lambda} - s_p\|_\rho = \|\sqrt{L_\mathcal{K}}(\hat{s}_{p,\lambda} - s_p)\|_{\mathcal{H}}$. Then we can apply the above discussion to obtain the bound for $\|\cdot\|_\rho$. \square

Next, we give the non-asymptotic version of Theorem 4.6 as follows

Theorem B.2. *Under the same assumption of Theorem B.1, we define $g_\lambda(\sigma) := (\lambda + \sigma)^{-1}$, and choose $\mathbf{Z} := \{\mathbf{z}^n\}_{n \in [N]} \subseteq \mathcal{X}$. Let $\mathbf{Y} := \{\mathbf{y}^m\}_{m \in [M]}$ be a set of i.i.d. samples drawn from ρ . Let $\hat{s}_{p,\lambda,\mathbf{Z}}$ be defined as in (8) with $\mathbf{X} = \mathbf{Z} \cup \mathbf{Y}$. Suppose $N = M^\alpha$, then for any $0 < \delta < 1$, $M \geq (2\sqrt{2}\kappa^2 \log(4/\delta))^{\frac{2r+2}{r}}$, choosing $\lambda = M^{-\frac{1}{2r+2}}$, the following inequalities hold with probability at least $1 - \delta$*

$$\sup_{\mathbf{Z}} \|\hat{s}_{p,\lambda,\mathbf{Z}} - s_p\|_{\mathcal{H}} \leq C_1 M^{-\frac{r}{2r+2}} \log \frac{4}{\delta} + C_3 M^{\alpha - \frac{r}{r+1}},$$

where $C_3 := 2(\kappa^2 + 1)^2 \|s_p\|_{\mathcal{H}}$, and the $\sup_{\mathbf{Z}}$ is taken over all $\{\mathbf{z}^n\}_{n \in [N]} \subset \mathcal{X}$.

In particular, when $\alpha = \frac{r}{2r+2}$, we have

$$\sup_{\mathbf{Z}} \|\hat{s}_{p,\lambda,\mathbf{Z}} - s_p\|_{\mathcal{H}} \leq (C_1 + C_3) M^{-\frac{r}{2r+2}} \log \frac{4}{\delta}.$$

Proof. We define $T_{\mathbf{Z}} := \frac{1}{N} S_{\mathbf{Z}}^* S_{\mathbf{Z}}$, where $S_{\mathbf{Z}} f := (f(\mathbf{z}^1), \dots, f(\mathbf{z}^N))$ is the sampling operator. Let $\hat{L}_\mathcal{K} := T_{\mathbf{Y}}$ and $\hat{s}_{p,\lambda}$ be the estimator obtained from \mathbf{Y} . Then we can write $\hat{s}_{p,\lambda,\mathbf{Z}} := g_\lambda(\hat{L}_\mathcal{K} + R_{\mathbf{Z}})(\hat{L}_\mathcal{K} + R_{\mathbf{Z}})s_p$, where $R_{\mathbf{Z}} := \frac{N}{M+N}(T_{\mathbf{Z}} - \hat{L}_\mathcal{K})$.

We can bound the error as follows

$$\begin{aligned} \|\hat{s}_{p,\lambda,\mathbf{Z}} - s_p\|_{\mathcal{H}} &\leq \|\hat{s}_{p,\lambda,\mathbf{Z}} - \hat{s}_{p,\lambda}\|_{\mathcal{H}} + \|\hat{s}_{p,\lambda} - s_p\|_{\mathcal{H}} \\ &\leq \|(g_\lambda(\hat{L}_\mathcal{K} + R_{\mathbf{Z}}) - g_\lambda(\hat{L}_\mathcal{K}))\hat{L}_\mathcal{K}s_p\|_{\mathcal{H}} + \|g_\lambda(\hat{L}_\mathcal{K} + R_{\mathbf{Z}})R_{\mathbf{Z}}s_p\|_{\mathcal{H}} + \|\hat{s}_{p,\lambda} - s_p\|_{\mathcal{H}}. \end{aligned}$$

The last term has been bounded by Theorem B.1, and we consider the first two terms. Since $g_\lambda(\sigma) = (\lambda + \sigma)^{-1}$ is Lipschitz in $[0, \infty)$, from Lemma D.5, we have $\|g_\lambda(\hat{L}_\mathcal{K} + R_{\mathbf{Z}}) - g_\lambda(\hat{L}_\mathcal{K})\|_{\text{HS}} \leq \|R_{\mathbf{Z}}\|_{\text{HS}}/\lambda^2$. Note $\|g_\lambda(\hat{L}_\mathcal{K} + R_{\mathbf{Z}})R_{\mathbf{Z}}\|_{\text{HS}} \leq \|R_{\mathbf{Z}}\|_{\text{HS}}/\lambda$, we obtain

$$\begin{aligned} \|\hat{s}_{p,\lambda,\mathbf{Z}} - \hat{s}_{p,\lambda}\|_{\mathcal{H}} &\leq \left(\frac{\kappa^2}{\lambda^2} + \frac{1}{\lambda} \right) \|R_{\mathbf{Z}}\|_{\text{HS}} \|s_p\|_{\mathcal{H}} \leq \left(\frac{\kappa^2}{\lambda^2} + \frac{1}{\lambda} \right) \frac{2\kappa^2 N}{M+N} \|s_p\|_{\mathcal{H}} \\ &\leq \frac{2(\kappa^2 + 1)^2 N}{\lambda^2 M} \|s_p\|_{\mathcal{H}} = 2(\kappa^2 + 1)^2 M^{\alpha - \frac{r}{r+1}} \|s_p\|_{\mathcal{H}}. \end{aligned}$$

Combining with Theorem B.1, and noticing that the right hand does not depend on \mathbf{Z} , we obtain the final bound. \square

Finally, we prove the error bound of the Stein estimator with its original out-of-sample extension.

Proof of Corollary 4.7. The Stein estimator at point $\mathbf{x} \in \mathcal{X}$ can be written as

$$\hat{s}_{p,\lambda,\mathbf{x}}(\mathbf{x}) = \sum_{i=1}^d \langle \mathcal{K}_{\mathbf{x}} \mathbf{e}_i, \hat{s}_{p,\lambda,\mathbf{x}} \rangle_{\mathcal{H}} \mathbf{e}_i,$$

where $\{\mathbf{e}_i\}$ is the standard basis of \mathbb{R}^d . Note that

$$\sup_{\mathbf{x} \in \mathcal{X}} \|\hat{s}_{p,\lambda,\mathbf{x}}(\mathbf{x}) - s_p(\mathbf{x})\|_2 \leq \sum_{i=1}^d \sup_{\mathbf{x} \in \mathcal{X}} |\langle \mathcal{K}_{\mathbf{x}} \mathbf{e}_i, \hat{s}_{p,\lambda,\mathbf{x}} - s_p \rangle_{\mathcal{H}}| \leq \kappa^2 \sup_{\mathbf{x} \in \mathcal{X}} \|\hat{s}_{p,\lambda,\mathbf{x}} - s_p\|_{\mathcal{H}}.$$

Then, the bound of Stein estimator immediately follows from Theorem 4.6. \square

C. Details in Section 3

C.1. A General Version of Nyström KEF

In this section, we briefly review the Nyström version of KEF (NKEF, Sutherland et al. (2018)) and give a more general version of it in our framework.

One of the drawbacks of KEF, as we have mentioned before, is the high computational complexity. It requires to solve an $Md \times Md$ linear system, where M is the sample size and d is the dimension. Note that the solution of KEF in (3) lies in the subspace generated by $\{\partial_i k(\mathbf{x}^m, \cdot) : i \in [d], m \in [M]\} \cup \{\hat{\zeta}\}$. The Nyström version of KEF consider to minimize the loss (2) in a smaller subspace generated by $\{\partial_i k(\mathbf{z}^n, \cdot) : i \in [d], n \in [N]\}$, where $N \ll M$ and $\{\mathbf{z}^n\}$ is a subset randomly sampled from $\{\mathbf{x}^m\}$. Sutherland et al. (2018) showed that it suffices to solve an $Nd \times Nd$ linear system, which reduces the computational complexity, while the convergence rate remains the same as that of KEF if $N = \Omega(M^\theta \log M)$, where $\theta \in [1/3, 1/2]$.

In our framework, we can also consider to find our estimator in a smaller subspace. Let $\mathcal{H}_{\mathbf{Z}}$ be the subspace generated by $\{\mathbf{z}^n\}_{n \in [N]}$, i.e., $\mathcal{H}_{\mathbf{Z}} := \text{span}\{\mathcal{K}_{\mathbf{z}^n} \mathbf{c} : n \in [N], \mathbf{c} \in \mathbb{R}^d\}$. Consider the minimization problem, which is a modification of (6), where the solution is found in $\mathcal{H}_{\mathbf{Z}}$:

$$\hat{s}_{p,\lambda}^{\mathbf{Z}} = \arg \min_{s \in \mathcal{H}_{\mathbf{Z}}} \frac{1}{M} \sum_{m=1}^M \|s(\mathbf{x}^m) - s_p(\mathbf{x}^m)\|_2^2 + \frac{\lambda}{2} \|s\|_{\mathcal{H}_{\mathcal{K}}}^2. \quad (16)$$

The solution can be written as $\hat{s}_{p,\lambda}^{\mathbf{Z}} = (P_{\mathbf{Z}} \hat{L}_{\mathcal{K}} P_{\mathbf{Z}} + \lambda I)^{-1} P_{\mathbf{Z}} \hat{\zeta}$, where $\hat{\zeta}, \hat{L}_{\mathcal{K}}$ are defined as in Sec. 3.1 and $P_{\mathbf{Z}} : \mathcal{H}_{\mathcal{K}} \rightarrow \mathcal{H}_{\mathcal{K}}$ is the projection operator onto $\mathcal{H}_{\mathbf{Z}}$, which can be defined as

$$P_{\mathbf{Z}} f := \arg \min_{g \in \mathcal{H}_{\mathbf{Z}}} \|g - f\|_{\mathcal{H}}^2 = S_{\mathbf{Z}}^* (S_{\mathbf{Z}} S_{\mathbf{Z}}^*)^{-1} S_{\mathbf{Z}} f,$$

where $S_{\mathbf{Z}}, S_{\mathbf{Z}}^*$ is the sampling operator and its adjoint, respectively. This motivates us to define the Nyström version of our score estimators for general regularization schemes as follows:

$$\hat{s}_{p,\lambda}^{g,\mathbf{Z}} := -g_{\lambda} (P_{\mathbf{Z}} \hat{L}_{\mathcal{K}} P_{\mathbf{Z}}) P_{\mathbf{Z}} \hat{\zeta}. \quad (17)$$

To obtain the matrix form of (17), we first introduce two operators:

$$\begin{aligned} \mathcal{L} &:= P_{\mathbf{Z}} \hat{L}_{\mathcal{K}} P_{\mathbf{Z}}, \\ \mathbf{L} &:= \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} \mathbf{K}_{\mathbf{Z}\mathbf{X}} \mathbf{K}_{\mathbf{X}\mathbf{Z}} \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}}. \end{aligned}$$

We want to connect the spectral decompositions of \mathbf{L} and \mathcal{L} as in Lemma 3.2. Suppose the spectral decomposition of \mathbf{L} is $\sum_{i=1}^{Md} \sigma_i \mathbf{u}_i \mathbf{u}_i^{\top}$, where $\|\mathbf{u}_i\|_{\mathbb{R}^{Md}} = 1$. Consider $v_i := S_{\mathbf{Z}}^* \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} \mathbf{u}_i$, we can verify that

$$\begin{aligned} \|v_i\|_{\mathcal{H}}^2 &= (\mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} \mathbf{u}_i)^{\top} \mathbf{K}_{\mathbf{Z}\mathbf{Z}} \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} \mathbf{u}_i = \mathbf{u}_i^{\top} \mathbf{u}_i = 1, \\ \mathcal{L} v_i &= S_{\mathbf{Z}}^* \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-1} \mathbf{K}_{\mathbf{Z}\mathbf{X}} \mathbf{K}_{\mathbf{X}\mathbf{Z}} \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} \mathbf{u}_i = S_{\mathbf{Z}}^* \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} \mathbf{L} \mathbf{u}_i = \sigma_i v_i. \end{aligned}$$

Thus, $\mathcal{L} = \sum_{i=1}^{Md} \sigma_i \langle v_i, \cdot \rangle_{\mathcal{H}} v_i$ is the spectral decomposition of \mathcal{L} . The estimator can be written as

$$\hat{s}_{p,\lambda}^{g,\mathbf{Z}} = -S_{\mathbf{Z}}^* \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} \left(\sum_{i=1}^{Md} g_\lambda(\sigma_i) \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} \mathbf{h} = -S_{\mathbf{Z}}^* \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} g_\lambda(\mathbf{L}) \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} \mathbf{h}. \quad (18)$$

The above estimator only involves smaller matrices. However, it requires some expensive matrix manipulations like the matrix square root for general regularization schemes. Fortunately, these expensive terms can be cancelled when using the Tikhonov regularization:

Example C.1. When we consider the Tikhonov regularization $g_\lambda(\sigma) = (\sigma + \lambda)^{-1}$ and curl-free kernels, the score estimator (18) becomes $\hat{s}_{p,\lambda}^{g,\mathbf{Z}}(\mathbf{x}) = -\mathbf{K}_{\mathbf{x}\mathbf{Z}} \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} (\mathbf{L} + \lambda \mathbf{I})^{-1} \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-\frac{1}{2}} \mathbf{h} = -\mathbf{K}_{\mathbf{x}\mathbf{Z}} (\mathbf{K}_{\mathbf{Z}\mathbf{X}} \mathbf{K}_{\mathbf{Z}\mathbf{Z}} + \lambda \mathbf{K}_{\mathbf{Z}\mathbf{Z}})^{-1} \mathbf{h}$. Similar to Example 3.5, we find this is exactly the same as the NKEF estimator obtained in Sutherland et al. (2018, Theorem 1).

C.2. Computational Details

Details of Example 3.6 Using the notation in Example 3.6 and Sec. 2.2, we can reformulate SSGE into a matrix form as follows:

$$\begin{aligned} \hat{g}_i(\mathbf{x}) &= - \sum_{j=1}^J \left(\frac{1}{M} \sum_{n=1}^M \partial_i \hat{\psi}_j(\mathbf{x}^n) \right) \psi_j(\mathbf{x}) \\ &= - \sum_{j=1}^J \frac{1}{M} \left(\frac{\sqrt{M}}{\lambda_j} \sum_{n,m=1}^M \partial_i k(\mathbf{x}^n, \mathbf{x}^m) w_j^{(m)} \right) \left(\frac{\sqrt{M}}{\lambda_j} \sum_{\ell=1}^M k(\mathbf{x}, \mathbf{x}^\ell) w_j^{(\ell)} \right) \\ &= - \sum_{j=1}^J \frac{1}{\lambda_j^2} \left(\sum_{n,m=1}^M \partial_i k(\mathbf{x}^n, \mathbf{x}^m) w_j^{(m)} \right) \left(\sum_{\ell=1}^M k(\mathbf{x}, \mathbf{x}^\ell) w_j^{(\ell)} \right) \\ &= - \sum_{\ell=1}^M k(\mathbf{x}, \mathbf{x}^\ell) \sum_{n,m=1}^M \left(\sum_{j=1}^J \frac{w_j^{(m)} w_j^{(\ell)}}{\lambda_j^2} \right) \partial_i k(\mathbf{x}^n, \mathbf{x}^m) \\ &= - \sum_{\ell=1}^M k(\mathbf{x}, \mathbf{x}^\ell) \sum_{m=1}^M \left(\sum_{j=1}^J \frac{w_j^{(m)} w_j^{(\ell)}}{\lambda_j^2} \right) \left(\sum_{n=1}^M \partial_i k(\mathbf{x}^n, \mathbf{x}^m) \right) \\ &= -k(\mathbf{x}, \mathbf{X}) \left(\sum_{j=1}^J \frac{\mathbf{w}_j \mathbf{w}_j^\top}{\lambda_j^2} \right) \mathbf{r}_i, \end{aligned}$$

where $r_{i,j} = \sum_{n=1}^M \partial_i k(\mathbf{x}^n, \mathbf{x}^j)$, and $\mathbf{w}_1, \dots, \mathbf{w}_M$ is the unit eigenvectors of $k(\mathbf{X}, \mathbf{X})$ corresponding to eigenvalues $\lambda_1 \geq \dots \geq \lambda_M$. $w_j^{(m)}$ is the m -th component of \mathbf{w}_j . Note that when using diagonal kernels, we have $\mathcal{K}(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y}) \otimes \mathbf{I}_d$, then the eigenvectors of $\mathcal{K}(\mathbf{X}, \mathbf{X})$ are $\{\mathbf{w}_i \otimes \mathbf{e}_j : i \in [M], j \in [d]\}$ and the eigenvalue corresponds to $\mathbf{w}_i \otimes \mathbf{e}_j$ is λ_i , where $\{\mathbf{e}_j\}$ is the standard basis of \mathbb{R}^d . We also note that in this case

$$h_{(m-1)d+i} = \hat{\zeta}(\mathbf{x}^m)_i = \frac{1}{M} \sum_{\ell=1}^M (\text{div}_{\mathbf{x}^\ell} \mathcal{K}(\mathbf{x}^\ell, \mathbf{x}^m))_i = \frac{1}{M} \sum_{\ell=1}^M \partial_i k(\mathbf{x}^\ell, \mathbf{x}^m) = M r_{i,m}.$$

Comparing with (12), we find that SSGE is equivalent to use diagonal kernels and spectral cut-off regularization.

Details of Example 3.7 For the regularizer $g_\lambda(\sigma) := (\lambda + \sigma)^{-1} \mathbf{1}_{\{\sigma > 0\}}$, from Lemma C.2 we know when \mathbf{K} is non-singular, $\hat{s}_{p,\lambda}^g(\mathbf{x}) = -\mathbf{K}_{\mathbf{x}\mathbf{X}} \mathbf{K}^{-1} (\frac{1}{M} \mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{h}$. Next, we consider the minimization problem in (6), and ignore the one-dimensional subspace $\mathbb{R} \hat{\zeta}$ of the solution space, and assume the solution is $\mathbf{K}_{\mathbf{x}\mathbf{X}} \mathbf{c}$ as before. We can rewrite the objective in (6) to

$$\frac{1}{M} \mathbf{c}^\top \mathbf{K}^2 \mathbf{c} + \lambda \mathbf{c} \mathbf{K} \mathbf{c} + 2 \mathbf{c}^\top \mathbf{h}.$$

By taking gradient, we find \mathbf{c} satisfies $(\frac{1}{M} \mathbf{K}^2 + \lambda \mathbf{K}) \mathbf{c} = -\mathbf{h}$, so it is equivalent to use the previously mentioned regularization.

C.3. Curl-Free Kernels

Recover the Function From Its Gradient. Since vector fields in a curl-free RKHS is always the gradient of some functions, it is possible to recover these functions from its gradient. Specifically, suppose the curl-free kernel is defined by $\mathcal{K}_{\text{cf}}(\mathbf{x}, \mathbf{y}) = -\nabla^2 \psi(\mathbf{x} - \mathbf{y})$ and $f \in \mathcal{H}_{\mathcal{K}_{\text{cf}}}$. Assume f is of the following form

$$f = \sum_{i=1}^m \mathcal{K}_{\text{cf}}(\mathbf{x}^i, \cdot) \mathbf{c}_i = - \sum_{i=1}^m \sum_{j=1}^d \nabla(\partial_j \psi(\mathbf{x}^i - \cdot)) c_i^{(j)} = \nabla \left(- \sum_{i=1}^m \sum_{j=1}^d \partial_j \psi(\mathbf{x}^i - \cdot) c_i^{(j)} \right),$$

where $c_i^{(j)}$ is the j -th component of \mathbf{c}_i . Then, we find a desired function whose gradient is f .

The Special Structure of $\mathcal{K}_{\text{cf}}(\mathbf{x}, \mathbf{y}) = -\nabla^2 \phi(\|\mathbf{x} - \mathbf{y}\|)$. As we have mentioned in Sec. 3.5, curl-free kernels have some special structures. Suppose \mathcal{K}_{cf} is a curl-free kernel defined by $\nabla^2 \phi(r)$, where $\mathbf{r} = (\mathbf{x} - \mathbf{x}')^T$ and $r = \|\mathbf{r}\|$. Then

$$\begin{aligned} \frac{\partial}{\partial r_i} \phi &= \phi' \frac{r_i}{r}, \\ \nabla \frac{\partial}{\partial r_i} \phi &= \phi'' \frac{r_i}{r^2} \mathbf{r} + \phi' \frac{\mathbf{e}_i r - r_i \frac{\mathbf{r}}{r}}{r^2}, \end{aligned}$$

where \mathbf{e}_i is the i -th column of the identity matrix. Then the curl-free kernel is of the form

$$\mathcal{K}_{\text{cf}}(\mathbf{x}, \mathbf{y}) = \left(\frac{\phi'}{r^3} - \frac{\phi''}{r^2} \right) \mathbf{r} \mathbf{r}^T - \frac{\phi'}{r} \mathbf{I}. \quad (19)$$

We also obtain a divergence formula for such kernel. Note that

$$\begin{aligned} \partial_{jj} \partial_i \phi &= \phi''' \frac{r_j^2 r_i}{r^3} + \phi'' \frac{(r_i + r_j \delta_{ij}) r^2 - 2r_j^2 r_i}{r^4} \\ &+ \phi'' \frac{r_j}{r} \frac{\delta_{ij} r - r_i \frac{r_j}{r}}{r^2} + \frac{\phi'}{r^6} [(\delta_{ij} r_j - r_i) r^3 - 3r r_j (\delta_{ij} r^2 - r_i r_j)], \end{aligned}$$

where $\delta_{ij} = [i = j]$. Next, we sum out j and then obtain

$$\text{div}_{\mathbf{x}} \mathcal{K}_{\text{cf}}(\mathbf{x}, \mathbf{x}') = -\Delta(\partial_i \phi)(r) = -\frac{\mathbf{r}}{r} \left[\phi'''(r) + \frac{d-1}{r} \left(\phi''(r) - \frac{\phi'(r)}{r} \right) \right]. \quad (20)$$

The Special Structure of $\mathcal{K}_{\text{cf}}(\mathbf{x}, \mathbf{y}) = -\nabla^2 \varphi(\|\mathbf{x} - \mathbf{y}\|^2)$. Since many frequently used kernels only depend on $\|\mathbf{x} - \mathbf{y}\|^2$, we consider the structure of curl-free kernels of these types. Suppose \mathcal{K}_{cf} is a curl-free kernel defined by $\nabla^2 \varphi(r^2)$, where $\mathbf{r} = (\mathbf{x} - \mathbf{x}')^T$ and $r = \|\mathbf{r}\|$. Then, using (19) and (20) we can find

$$\mathcal{K}_{\text{cf}}(\mathbf{x}, \mathbf{y}) = -4\varphi'' \mathbf{r} \mathbf{r}^T - 2\varphi' \mathbf{I}, \quad (21)$$

$$\text{div}_{\mathbf{x}} \mathcal{K}_{\text{cf}}(\mathbf{x}, \mathbf{y}) = -4[(d+2)\varphi'' + 2r^2 \varphi'''] \mathbf{r}. \quad (22)$$

C.4. Details of Different Regularization Schemes

C.4.1. TIKHONOV REGULARIZATION

Proof of Theorem 3.1. When $g_\lambda(\sigma) = (\sigma + \lambda)^{-1}$, the estimator is $\hat{s}_{p,\lambda} = -(\hat{L}_{\mathcal{K}} + \lambda I)^{-1} \hat{\zeta}$. We need to compute the explicit formula of the inverse of $\hat{L}_{\mathcal{K}} + \lambda I$. Note that $(\hat{L}_{\mathcal{K}} + \lambda I)^{-1} \hat{\zeta}$ is the solution of the following minimization problem

$$\hat{s}_{p,\lambda}^g = \arg \min_{s \in \mathcal{H}_{\mathcal{K}}} \frac{1}{M} \sum_{i=1}^M s(\mathbf{x}^i)^T s(\mathbf{x}^i) + 2\langle s, \hat{\zeta} \rangle_{\mathcal{H}} + \lambda \|s\|_{\mathcal{H}}^2.$$

From the general representer theorem (Sriperumbudur et al., 2017, Theorem A.2), the minimizer lies in the space generated by

$$\{\mathcal{K}_{\mathbf{x}^i} \mathbf{c} : i \in [M], \mathbf{c} \in \mathbb{R}^d\} \cup \{\hat{\zeta}\}.$$

We can assume

$$\hat{s}_{p,\lambda}^g = \sum_{i=1}^M \mathcal{K}_{\mathbf{x}^i} \mathbf{c}_i + a \hat{\zeta}.$$

Define $\mathbf{c} := (\mathbf{c}_1, \dots, \mathbf{c}_M)$ and $\mathbf{h} := (\hat{\zeta}(\mathbf{x}^1), \dots, \hat{\zeta}(\mathbf{x}^M))$, then the optimization objective can be written as

$$\frac{1}{M} (\mathbf{c}^\top \mathbf{K}^2 \mathbf{c} + 2a \mathbf{c}^\top \mathbf{K} \mathbf{h} + a^2 \mathbf{h}^\top \mathbf{h}) + 2(a \|\hat{\zeta}\|_{\mathcal{H}}^2 + \mathbf{h}^\top \mathbf{c}) + \lambda (\mathbf{c}^\top \mathbf{K} \mathbf{c} + 2a \mathbf{c}^\top \mathbf{h} + a^2 \|\hat{\zeta}\|_{\mathcal{H}}^2).$$

Taking the derivative, we need to solve the following linear system

$$\begin{aligned} \frac{1}{M} (\mathbf{K}^2 \mathbf{c} + a \mathbf{K} \mathbf{h}) + \mathbf{h} + \lambda (\mathbf{K} \mathbf{c} + a \mathbf{h}) &= 0, \\ \frac{1}{M} (a \mathbf{h}^\top \mathbf{h} + \mathbf{c}^\top \mathbf{K} \mathbf{h}) + (1 + \lambda a) \|\hat{\zeta}\|_{\mathcal{H}}^2 + \lambda \mathbf{c}^\top \mathbf{h} &= 0. \end{aligned}$$

By some calculations, this system is equivalent to $a = -1/\lambda$ and $(\mathbf{K} + M\lambda I)\mathbf{c} = \mathbf{h}/\lambda$. □

C.4.2. SPECTRAL CUT-OFF REGULARIZATION

Proof of Lemma 3.2. Let \mathcal{H}_0 be the subspace of $\mathcal{H}_{\mathcal{K}}$ generated by $\{\mathcal{K}_{\mathbf{x}^m} \mathbf{c} : \mathbf{c} \in \mathbb{R}^d, m \in [M]\}$. Note that $f(\mathbf{x}^m)^\top \mathbf{c} = \langle \mathcal{K}(\cdot, \mathbf{x}^m) \mathbf{c}, f \rangle_{\mathcal{H}} = 0$ for any $f \in \mathcal{H}_0^\perp$ and $\mathbf{c} \in \mathbb{R}^d$. We know $\hat{L}_{\mathcal{K}} = 0$ on \mathcal{H}_0^\perp . Also note $\hat{L}_{\mathcal{K}} v \in \mathcal{H}_0$ and $v(\mathbf{x}^m) = \mathbf{u}^{(m)} \sqrt{M\sigma}$, then

$$\hat{L}_{\mathcal{K}} v(\mathbf{x}^k) = \frac{1}{M} \sum_{m=1}^M \mathcal{K}(\mathbf{x}^k, \mathbf{x}^m) v(\mathbf{x}^m) = \frac{1}{\sqrt{M}} \sum_{m=1}^M \mathcal{K}(\mathbf{x}^k, \mathbf{x}^m) \sqrt{\sigma} \mathbf{u}^{(m)} = \sigma v(\mathbf{x}^k),$$

and we conclude that $\hat{L}_{\mathcal{K}} v = \sigma v$. The following equation shows v is normalized:

$$\|v\|_{\mathcal{H}}^2 = \frac{1}{\sqrt{M\sigma}} \sum_{m=1}^M \langle \mathcal{K}(\cdot, \mathbf{x}^m) \mathbf{u}^{(m)}, v \rangle_{\mathcal{H}} = \frac{1}{\sqrt{M\sigma}} \sum_{m=1}^M \langle \mathbf{u}^{(m)}, v(\mathbf{x}^m) \rangle_{\mathbb{R}^d} = \sum_{m=1}^M (\mathbf{u}^{(m)})^\top \mathbf{u}^{(m)} = 1.$$

□

Theorem 3.3 is a corollary of the following lemma, which provides a general form for the regularizer g_λ with $g_\lambda(0) = 0$.

Lemma C.2. *Let $g_\lambda : [0, \kappa^2] \rightarrow \mathbb{R}$ be a regularizer such that $g_\lambda(0) = 0$. Let $(\sigma_j, \mathbf{u}_j)_{j \geq 1}$ be the non-zero eigenvalue and eigenvector pairs that satisfy $\frac{1}{M} \mathbf{K} \mathbf{u}_j = \sigma_j \mathbf{u}_j$. Then we have*

$$g_\lambda(\hat{L}_{\mathcal{K}}) \hat{\zeta} = \mathbf{K}_{\mathbf{x}\mathbf{x}} \left(\sum \frac{g_\lambda(\sigma_i)}{M\sigma_i} \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{h},$$

where $\mathbf{K}_{\mathbf{x}\mathbf{x}}$ and \mathbf{h} are defined as in Theorem 3.1.

Proof. Let $\{(\mu_i, v_i)\}$ be the pairs of non-zero eigenvalues and eigenfunctions of $\hat{L}_{\mathcal{K}} : \mathcal{H} \rightarrow \mathcal{H}$, then by Lemma 3.2 we have $\sigma_i = \mu_i$. Note that

$$\hat{L}_{\mathcal{K}} = \sum \mu_i \langle v_i, \cdot \rangle_{\mathcal{H}} v_i \quad \text{and} \quad g_\lambda(\hat{L}_{\mathcal{K}}) = \sum g_\lambda(\mu_i) \langle v_i, \cdot \rangle_{\mathcal{H}} v_i.$$

From Lemma 3.2, we have

$$\begin{aligned}
 g_\lambda(\hat{L}_\mathcal{K})\hat{\zeta} &= \sum g_\lambda(\sigma_i) \langle v_i, \hat{\zeta} \rangle_{\mathcal{H}} v_i \\
 &= \sum \left\{ g_\lambda(\sigma_i) \left\langle \frac{1}{\sqrt{M}\sigma_i} \sum_{j=1}^M \mathcal{K}_{\mathbf{x}^j} \mathbf{u}_i^{(j)}, \hat{\zeta} \right\rangle_{\mathcal{H}} \frac{1}{\sqrt{M}\sigma_i} \sum_{k=1}^M \mathcal{K}_{\mathbf{x}^k} \mathbf{u}_i^{(k)} \right\} \\
 &= \frac{1}{M} \sum \sum_{j,k=1}^M g_\lambda(\sigma_i) \sigma_i^{-1} \left\langle \mathcal{K}_{\mathbf{x}^j} \mathbf{u}_i^{(j)}, \hat{\zeta} \right\rangle_{\mathcal{H}} \mathcal{K}_{\mathbf{x}^k} \mathbf{u}_i^{(k)} \\
 &= \frac{1}{M} \sum \sum_{j,k=1}^M g_\lambda(\sigma_i) \sigma_i^{-1} \hat{\zeta}(\mathbf{x}^j)^\top \mathbf{u}_i^{(j)} \mathcal{K}_{\mathbf{x}^k} \mathbf{u}_i^{(k)} \\
 &= \mathcal{K}_{\mathbf{X}\mathbf{X}} \left(\sum \frac{g_\lambda(\sigma_i)}{M\sigma_i} \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{h}.
 \end{aligned}$$

□

C.4.3. ITERATIVE REGULARIZATION

Theorem C.3 (Landweber iteration). *Let $\hat{s}_{p,\lambda}^g$ be defined as in (8), and $g_\lambda(\sigma) = \eta \sum_{i=0}^{t-1} (1 - \eta\sigma)^i$, where $t := \lfloor \lambda^{-1} \rfloor$. Then we have*

$$\hat{s}_{p,\lambda}^g(\mathbf{x}) = -t\eta\hat{\zeta}(\mathbf{x}) + \mathbf{K}_{\mathbf{X}\mathbf{X}}\mathbf{c}_t,$$

where $\mathbf{c}_0 = 0$ and $\mathbf{c}_{t+1} = (\mathbf{I}_d - \eta\mathbf{K}/M)\mathbf{c}_t - t\eta^2\mathbf{h}/M$, and $\mathbf{K}_{\mathbf{X}\mathbf{X}}$ and \mathbf{h} are defined as in Theorem 3.1.

Proof. We note that the iteration process is

$$\begin{aligned}
 \hat{s}_p^{(1)} &= -\eta\hat{\zeta}, \\
 \hat{s}_p^{(t)} &= -\eta\hat{\zeta} + (I - \eta\hat{L}_\mathcal{K})\hat{s}_p^{(t-1)} \\
 &= \hat{s}_p^{(t-1)} + \eta(-\hat{\zeta} - \hat{L}_\mathcal{K}\hat{s}_p^{(t-1)}).
 \end{aligned}$$

where we define $\hat{s}_p^{(t)} := \hat{s}_{p,1/t}$. We can assume

$$\hat{s}_p^{(t)} = a_t\hat{\zeta} + \mathbf{K}_{\mathbf{X}\mathbf{X}}\mathbf{c}_t.$$

Then, by induction,

$$\begin{aligned}
 \hat{s}_p^{(t)} &= -\eta\hat{\zeta} + (I - \eta\hat{L}_\mathcal{K})(a_{t-1}\hat{\zeta} + \mathbf{K}_{\mathbf{X}\mathbf{X}}\mathbf{c}_{t-1}) \\
 &= (a_{t-1} - \eta)\hat{\zeta} + \mathbf{K}_{\mathbf{X}\mathbf{X}}(\mathbf{c}_{t-1} + \eta a_{t-1}\mathbf{h}/M - \eta\mathbf{K}\mathbf{c}_{t-1}/M).
 \end{aligned}$$

Thus, we have $a_t = -t\eta$ and $\mathbf{c}_t = (\mathbf{I} - \eta\mathbf{K}/M)\mathbf{c}_{t-1} - (t-1)\eta^2\mathbf{h}/M$, and $\mathbf{c}_1 = 0$. □

Before introducing the ν -method, we recall that the iterative regularization can be represented by a family of polynomials $g_\lambda(\sigma) = \text{poly}(\sigma)$, where g_λ converges to the function $1/\sigma$ as $\lambda \rightarrow 0$. For example, in the Landweber iteration we see that

$$g_\lambda(\sigma) = \eta \sum_{i=0}^{t-1} (1 - \eta\sigma)^i = \frac{1 - (1 - \eta\sigma)^t}{\sigma}.$$

We can verify that the identification of λ and t^{-1} satisfies Definition 4.1 about the regularization. To see the qualification, we note that the maximum $|1 - \sigma g_\lambda(\sigma)|\sigma^r = \sigma^r(1 - \eta\sigma)^t$ over $[0, \eta^{-1}]$ is attained when $\sigma = r/(r\eta + t)$ and hence

$$\sup_{0 \leq \sigma \leq \eta^{-1}} |1 - \sigma g_\lambda(\sigma)|\sigma^r \leq \frac{t^t r^r}{(r\eta + t)^{r+t}} \leq \left(\frac{r}{t}\right)^r = \max(r^r, 1)\lambda^r.$$

Thus, we see that the qualification is ∞ .

Example C.4 (ν -method). The ν -method (Engl et al., 1996) is an accelerated version of the Landweber iteration. The idea behind it is to find better polynomials $p_t(\sigma)$ to approximate the function $1/\sigma$, where p_t is a polynomial of degree t . These polynomials satisfy $\sup_{0 \leq \sigma \leq 1} |1 - \sigma p_t(\sigma)| \sigma^\nu \leq c_\nu t^{2\nu}$. Compared with the definition of the qualification in Definition 4.1, we can identify λ and t^{-2} . Thus, for the same regularization parameter, the ν -method only requires about $\lambda^{-1/2}$ iterations while the Landweber iteration requires about λ^{-1} iterations. For more details about the construction of these polynomials, we refer the readers to Engl et al. (1996, Appendix A.1 and Section 6.3)

Below we give the algorithm of the ν -method, where $t = \lfloor \lambda^{-1/2} \rfloor$ and $\hat{s}_{p,\lambda} := \hat{s}_p^{(t)}$.

$$\begin{aligned} \hat{s}_p^{(0)} &= 0, \quad \hat{s}_p^{(1)} = -\omega_1 \hat{\zeta}, \\ \hat{s}_p^{(t)} &= \hat{s}_p^{(t-1)} + u_t (\hat{s}_p^{(t-1)} - \hat{s}_p^{(t-2)}) + \omega_t (-\hat{\zeta} - \hat{L}_{\mathcal{K}} \hat{s}_p^{(t-1)}), \end{aligned}$$

where

$$\begin{aligned} u_t &= \frac{(t-1)(2t-3)(2t+2\nu-1)}{(t+2\nu-1)(2t+4\nu-1)(2t+2\nu-3)}, \\ \omega_t &= \frac{4(2t+2\nu-1)(t+\nu-1)}{(t+2\nu-1)(2t+4\nu-1)}. \end{aligned}$$

Similarly, we can assume

$$\hat{s}_p^{(t)} = a_t \hat{\zeta} + \mathbf{K}_{\mathbf{X}\mathbf{X}} \mathbf{c}_t.$$

Then, by induction,

$$\begin{aligned} \hat{s}_p^{(t)} &= \left(1 + u_t - \omega_t \hat{L}_{\mathcal{K}}\right) \hat{s}_p^{(t-1)} - u_t \hat{s}_p^{(t-2)} - \omega_t \hat{\zeta} \\ &= \left(1 + u_t - \omega_t \hat{L}_{\mathcal{K}}\right) (a_{t-1} \hat{\zeta} + \mathbf{K}_{\mathbf{X}\mathbf{X}} \mathbf{c}_{t-1}) - u_t (a_{t-2} \hat{\zeta} + \mathbf{K}_{\mathbf{X}\mathbf{X}} \mathbf{c}_{t-2}) - \omega_t \hat{\zeta} \\ &= ((1 + u_t) a_{t-1} - u_t a_{t-2} - \omega_t) \hat{\zeta} \\ &\quad + \mathbf{K}_{\mathbf{X}\mathbf{X}} \left((1 + u_t) \mathbf{c}_{t-1} - \frac{\omega_t}{M} (a_{t-1} \mathbf{h} + \mathbf{K} \mathbf{c}_{t-1}) - u_t \mathbf{c}_{t-2} \right). \end{aligned}$$

Thus, we obtain the iteration formula for a_t and \mathbf{c}_t as follows:

$$\begin{aligned} a_t &:= (1 + u_t) a_{t-1} - u_t a_{t-2} - \omega_t, \\ \mathbf{c}_t &:= (1 + u_t) \mathbf{c}_{t-1} - \frac{\omega_t}{M} (a_{t-1} \mathbf{h} + \mathbf{K} \mathbf{c}_{t-1}) - u_t \mathbf{c}_{t-2}, \end{aligned}$$

and $\mathbf{c}_0 = \mathbf{c}_1 = 0$, $a_0 = 0$, $a_1 = -\omega_1$.

D. Technical Results

Lemma D.1. Suppose Assumption B.5 holds, then $L_{\mathcal{K}}, \hat{L}_{\mathcal{K}} : \mathcal{H}_{\mathcal{K}} \rightarrow \mathcal{H}_{\mathcal{K}}$ are positive, self-adjoint, trace class operators. Moreover, $\text{tr } L_{\mathcal{K}} \leq \kappa^2$ and $\text{tr } \hat{L}_{\mathcal{K}} \leq \kappa^2$.

Proof. The result follows from a simple calculation. It is easy to see $L_{\mathcal{K}}$ and $\hat{L}_{\mathcal{K}}$ are positive and self-adjoint. We prove they are in trace class. Let $\{\varphi_i\}$ be a orthonormal basis of $\mathcal{H}_{\mathcal{K}}$ and $\{\mathbf{e}_i\}$ be the standard basis of \mathbb{R}^d , then

$$\begin{aligned} \text{tr } L_{\mathcal{K}} &= \sum_i \langle L_{\mathcal{K}} \varphi_i, \varphi_i \rangle_{\mathcal{H}} = \int_{\mathcal{X}} \sum_i \langle \mathcal{K}_{\mathbf{x}} \varphi_i, \varphi_i \rangle_{\mathcal{H}} d\rho = \sum_{k=1}^d \int_{\mathcal{X}} \sum_i \langle \langle \mathcal{K}_{\mathbf{x}} \mathbf{e}_k, \varphi_i \rangle_{\mathcal{H}} \mathcal{K}_{\mathbf{x}} \mathbf{e}_k, \varphi_i \rangle_{\mathcal{H}} d\rho \\ &= \sum_{k=1}^d \int_{\mathcal{X}} \sum_i |\langle \mathcal{K}_{\mathbf{x}} \mathbf{e}_k, \varphi_i \rangle_{\mathcal{H}}|^2 d\rho = \sum_{k=1}^d \int_{\mathcal{X}} \|\mathcal{K}_{\mathbf{x}} \mathbf{e}_k\|_{\mathcal{H}}^2 d\rho = \int_{\mathcal{X}} \text{tr } \mathcal{K}(\mathbf{x}, \mathbf{x}) d\rho \leq \kappa^2 \end{aligned}$$

Similarly, we have $\text{tr } \hat{L}_{\mathcal{K}} \leq \kappa^2$. □

We need the following concentration inequality in Hilbert spaces used in [Bauer et al. \(2007\)](#).

Lemma D.2 ([Bauer et al. \(2007\)](#), Proposition 23). *Let ξ be a random variable with values in a real Hilbert space H . Assume there are two constants σ, H , such that*

$$\mathbb{E}[\|\xi - \mathbb{E}\xi\|_H^m] \leq \frac{1}{2}m!\sigma^2 H^{m-2}, \quad \forall m \geq 2.$$

Then, for all $n \in \mathbb{N}$, $0 < \delta < 1$, the following inequality holds with probability at least $1 - \delta$

$$\|\hat{\xi} - \mathbb{E}\xi\|_H \leq 2 \left(\frac{H}{n} + \frac{\sigma}{\sqrt{n}} \right) \log \frac{2}{\delta},$$

where $\hat{\xi} = \frac{1}{n} \sum_{i=1}^n \xi_i$ and $\{\xi_i\}$ are independent copies of ξ .

Lemma D.3. *Under Assumption B.4, we have for all $M \in \mathbb{N}$, $0 < \delta < 1$, the following inequality holds with probability at least $1 - \delta$*

$$\|\hat{\zeta} - \zeta\|_{\mathcal{H}} \leq 2 \left(\frac{K}{M} + \frac{\Sigma}{\sqrt{M}} \right) \log \frac{2}{\delta}, \quad (23)$$

where $\hat{\zeta} = \frac{1}{M} \sum_{m=1}^M \text{div}_{\mathbf{x}^m} \mathcal{K}_{\mathbf{x}^m}^{\top}$ and $\{\mathbf{x}^m\}$ is the set of i.i.d. samples from ρ .

Proof. Define an $\mathcal{H}_{\mathcal{K}}$ -valued random variable $\xi_{\mathbf{x}} := \text{div}_{\mathbf{x}} \mathcal{K}_{\mathbf{x}}^{\top}$. It is easy to see $\mathbb{E}_{\mathbf{x} \sim \nu}[\xi_{\mathbf{x}}] = -L_{\mathcal{K}S_p} =: \xi$. From Assumption B.4, we have for $m \geq 2$,

$$\mathbb{E}_{\nu}[\|\xi_{\mathbf{x}} - \xi\|_{\mathcal{H}}^m] \leq m!K^m \mathbb{E}_{\nu} \left[\exp \left(\frac{\|\xi_{\mathbf{x}} - \xi\|_{\mathcal{H}}}{K} \right) - \frac{\|\xi_{\mathbf{x}} - \xi\|_{\mathcal{H}}}{K} - 1 \right] \leq \frac{1}{2}m!\Sigma^2 K^{m-2}.$$

Note that $\hat{\zeta} = \frac{1}{M} \sum_{m=1}^M \xi_{\mathbf{x}^m}$ and $\mathbb{E}_{\nu} \hat{\zeta} = \xi$. Then (23) follows from Lemma D.2. □

Lemma D.4. *Under Assumption B.5, we have for all $M \in \mathbb{N}$, $0 < \delta < 1$, the following inequality holds with probability at least $1 - \delta$*

$$\|\hat{L}_{\mathcal{K}} - L_{\mathcal{K}}\|_{\mathcal{H}} \leq \frac{2\sqrt{2}\kappa^2}{\sqrt{M}} \sqrt{\log \frac{2}{\delta}}. \quad (24)$$

Proof. This is a direct consequence of [Vito et al. \(2005, Lemma 8\)](#) and Lemma D.1. □

The following useful lemma is from [De Vito et al. \(2014, Lemma 7\)](#) and [Sriperumbudur et al. \(2017, Lemma 15\)](#)

Lemma D.5. *Suppose S and T are two self-adjoint Hilbert-Schmidt operators on a separable Hilbert space H with spectrum contained in the interval $[a, b]$. Given a Lipschitz function $r : [a, b] \rightarrow \mathbb{R}$ with Lipschitz constant L_r , we have*

$$\|r(S) - r(T)\|_{\text{HS}} \leq L_r \|S - T\|_{\text{HS}}.$$

E. Samples

Table 5. WAE samples on MNIST.



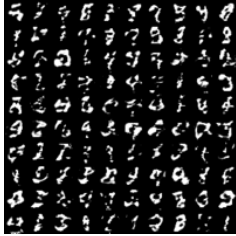





















	$d = 8$	$d = 32$	$d = 64$	$d = 128$
Stein				
SSGE				
SSM				
NKEF ₂				
ν -method				
KEF-CG				

Table 6. WAE samples on CelebA.

