THE UNIVERSITY OF TEXAS AT AUSTIN

pSVGD: Projected Stein Variational Gradient Descent

A fast and scalable Bayesian inference method in high dimensions

Peng Chen and Omar Ghattas {peng, omar}@oden.utexas.edu

INSTITUTE FOR COMPUTATIONAL ENGINEERING & SCIENCES — ODEN

Abstract

The **curse of dimensionality** is a longstanding challenge in Bayesian inference in high dimensions. In particular the Stein variational gradient descent method (SVGD) suffers from this curse. To overcome this challenge, we propose the projected SVGD (pSVGD) method, which exploits the **intrinsic low dimensionality** of the **data informed subspace** stemming from ill-posedness of inference problems. We adaptively construct the subspace using a **gradient information matrix** of the log-likelihood, and apply pSVGD to the much lower-dimensional coefficients of the parameter projection. The method is demonstrated to be **more accurate and efficient** than SVGD, and **scalable** with respect to the number of **parameters, samples, data points, and processor cores.**

Bayesian Inference and SVGD

Bayes' rule for parameter $x \in \mathbb{R}^d$

$$\underbrace{p(x)}_{\text{posterior}} = \frac{1}{Z} \underbrace{f(x)}_{\text{likelihood}} \underbrace{p_0(x)}_{\text{prior}} .$$

SVGD update for prior samples $x_1^0, \dots, x_N^0$

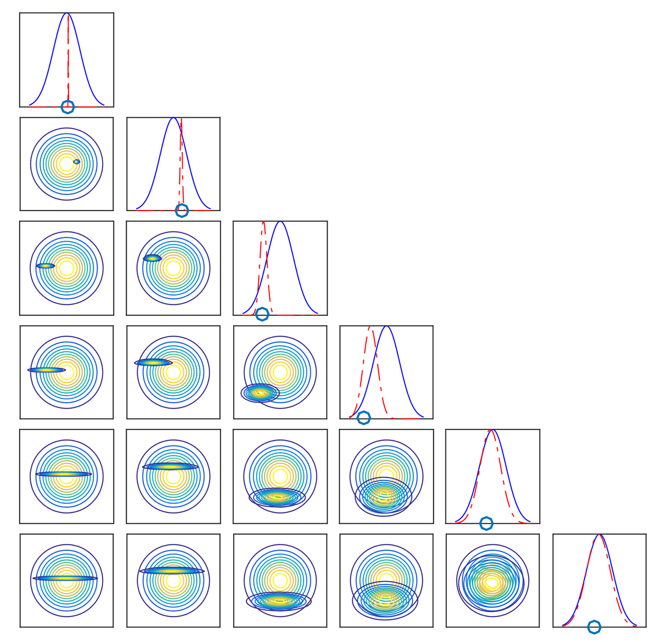
$$x_m^{\ell+1} = x_m^\ell + \epsilon_l \hat{\phi}_\ell^*(x_m^\ell), \quad m = 1,\dots,N, \ell = 0,1,\dots,$$

where $\hat{\phi}_\ell^*(x_m^\ell)$ is the approximate steepest direction

$$\hat{\phi}_\ell^*(x_m^\ell) = \frac{1}{N}\sum_{n=1}^N \nabla_{x_n^\ell} \log p(x_n^\ell) k(x_n^\ell, x_m^\ell) + \nabla_{x_n^\ell} k(x_n^\ell, x_m^\ell),$$

where $k(\cdot,\cdot)$ is, e.g., Gaussian kernel, collapsing in $\mathbb{R}^d$ for big d .

Data-informed Intrinsic Low-dimensionality



Gradient information matrix of the log-likelihood

$$H = \int_{\mathbb{R}^d} (\nabla_x \log f(x))(\nabla_x \log f(x))^T p(x) dx$$
$$\approx \sum_{m=1}^N \nabla_x \log f(x_m)(\nabla_x \log f(x_m))^T,$$

where $x_1, \dots, x_N$ are adaptively updated at suitable iteration.

Given prior covariance Γ, solve generalized eigenvalue problem:

$$H\psi_i = \lambda_i \Gamma \psi_i,$$

with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq \cdots \geq 0.$

Dimension Reduction by Projection

We project the high-dimensional pamameter to the subspace $X_r = \text{span}(\psi_1, \dots, \psi_r)$

$$P_r x := \sum_{i=1}^r \psi_i \psi_i^T x = \Psi_r w, \quad \forall x \in \mathbb{R}^d,$$

and define a projected posterior with profile function $g(P_r x)$

$$p_r(x) := \frac{1}{Z_r} g(P_r x) p_0(x) .$$

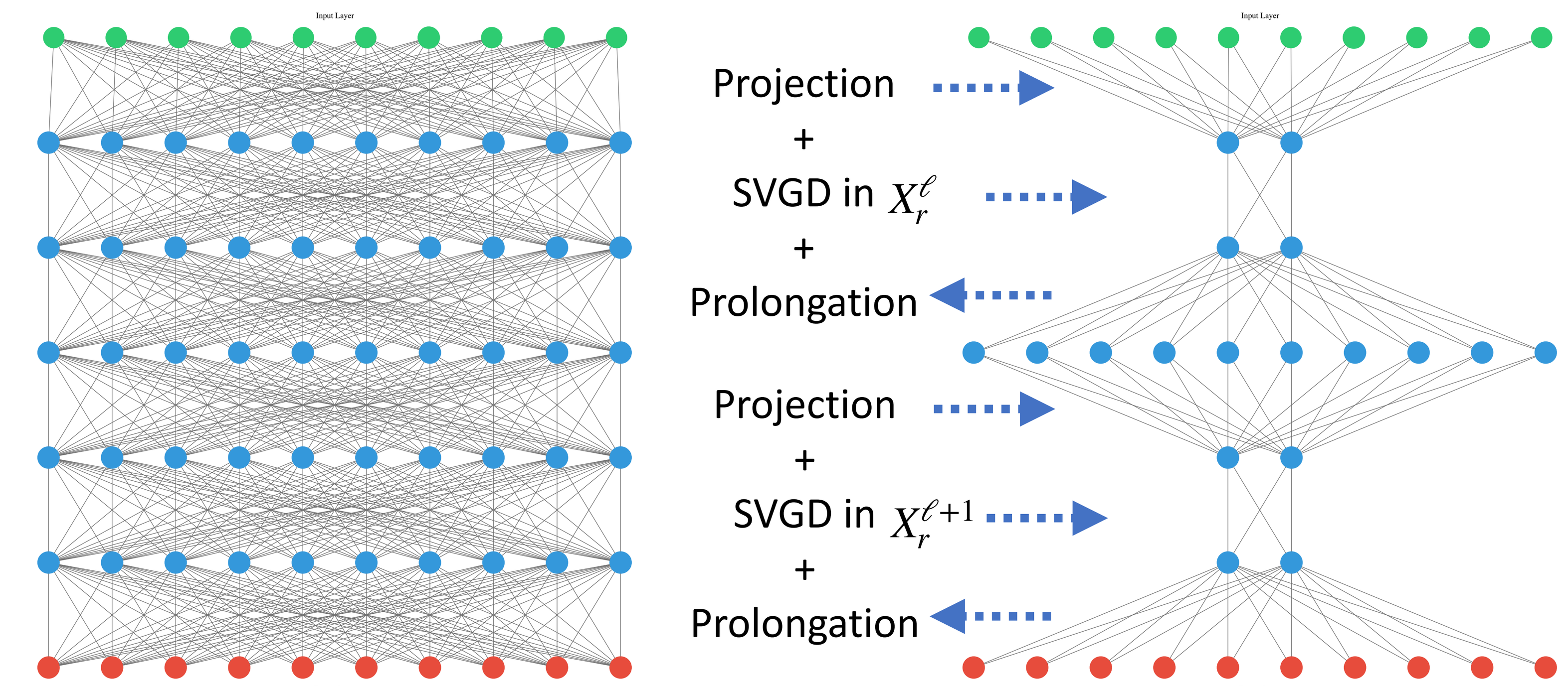
Optimal profile function in Kullback—Leibler divergence bounded by eigenvalues

$$g^*(P_r x) = \int_{X_\perp} f(P_r x + \xi) p_0^\perp(\xi | P_r x) d\xi \Longrightarrow D_{\text{KL}}(p|p_r) \geq D_{\text{KL}}(p|p_r^*) \leq \frac{\gamma}{2} \sum_{i=r+1}^d \lambda_i,$$

where the conditional/marginal density in complement subspace $X_\perp$

$$p_0^\perp(\xi | P_r x) = p_0(P_r x + \xi)/p_0^r(P_r x) \text{ with } p_0^r(P_r x) = \int_{X_\perp} p_0(P_r x + \xi) d\xi .$$

SVGD vs pSVGD



Bayes' rule for projection coefficient $w \in \mathbb{R}^r$ in data-informed dominant subspaces

$$\underbrace{\pi(w)}_{\text{posterior}} = \frac{1}{Z_w} \underbrace{g(\Psi_r w)}_{\text{likelihood}} \underbrace{\pi_0(w)}_{\text{prior}}$$

SVGD update for prior samples $w_1^0, \dots, w_N^0$

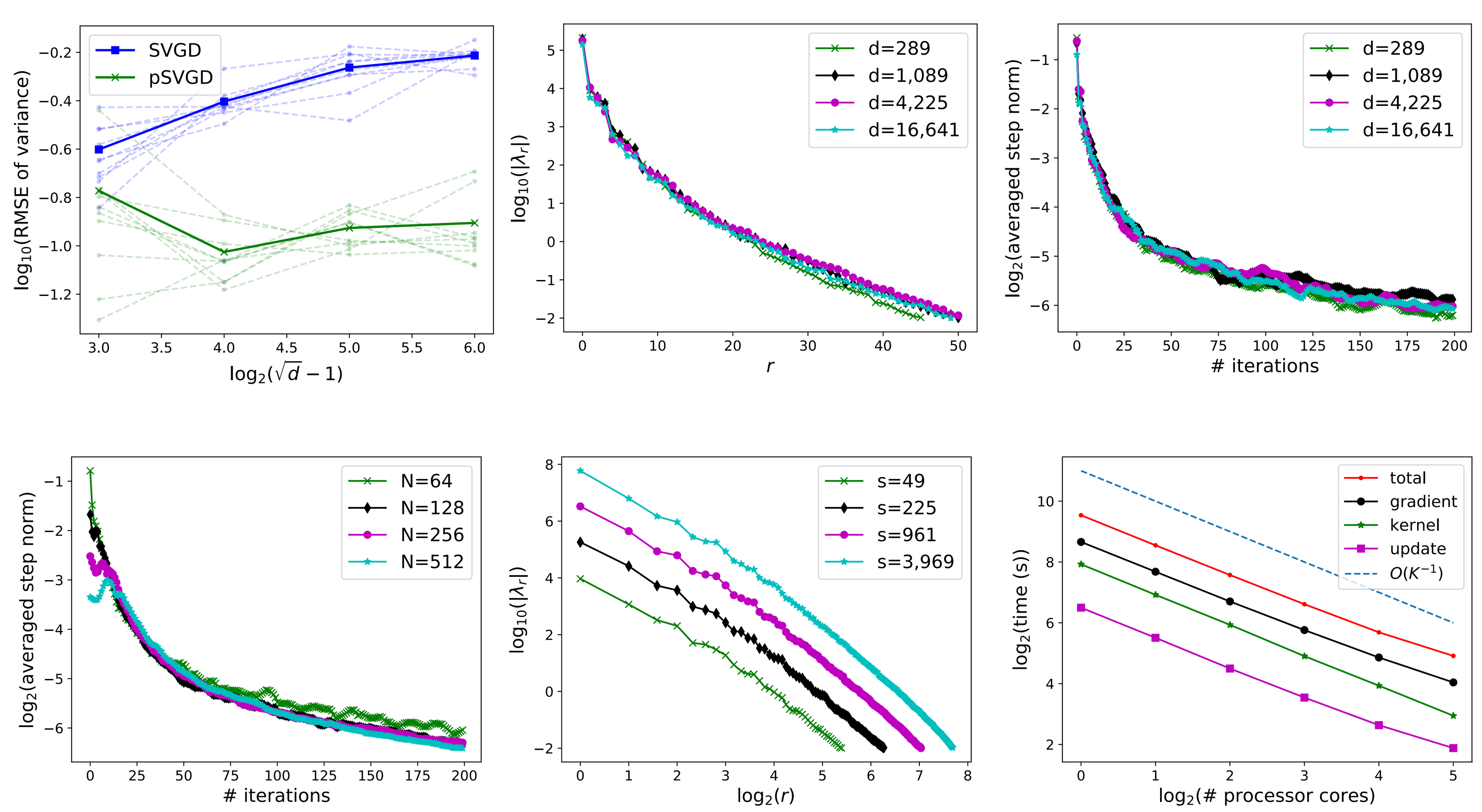
$$w_m^{\ell+1} = w_m^\ell + \epsilon_l \hat{\phi}_\ell^*(w_m^\ell), \quad m = 1,\dots,N, \ell = 0,1,\dots,$$

where $\hat{\phi}_\ell^*(w_m^\ell)$ is the approximate steepest direction

$$\hat{\phi}_\ell^*(w_m^\ell) = \frac{1}{N}\sum_{n=1}^N \nabla_{w_n^\ell} \log \pi(w_n^\ell) k_r(w_n^\ell, w_m^\ell) + \nabla_{w_n^\ell} k_r(w_n^\ell, w_m^\ell),$$

where $k_r(\cdot,\cdot)$ is the projected kernel in $\mathbb{R}^r$, $\nabla_w \log \pi(w) = \Psi_r^T \nabla_x \log p_r(P_r x)$.

Experiment I: PDE-constrained Inference

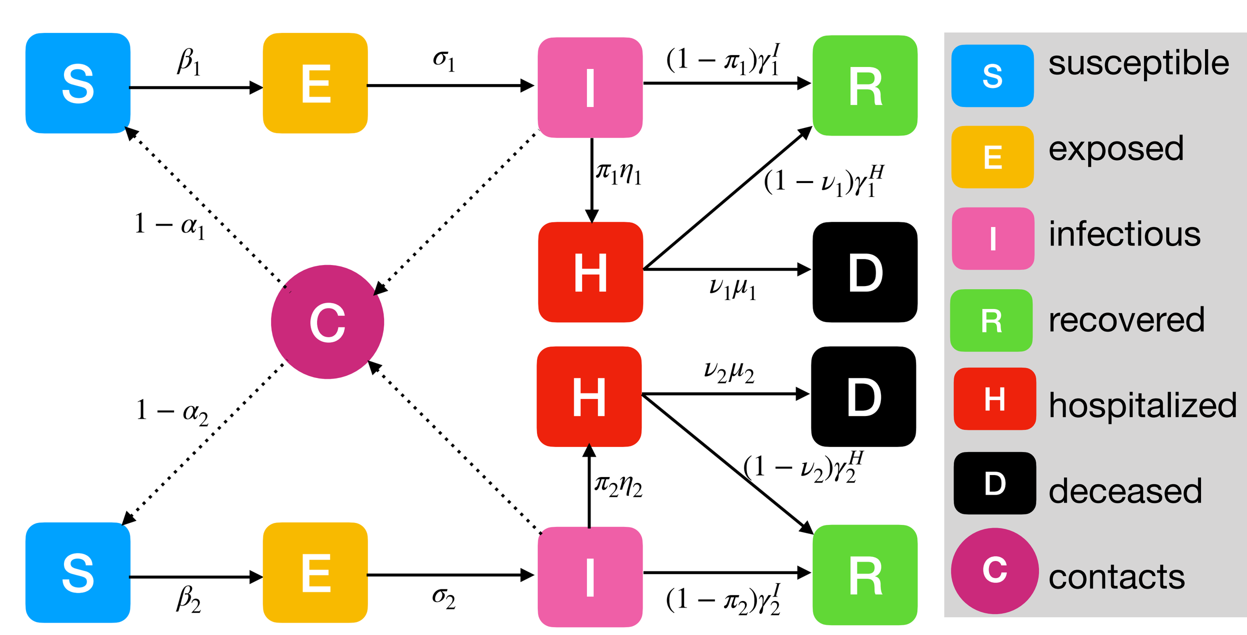


PDE model with random field

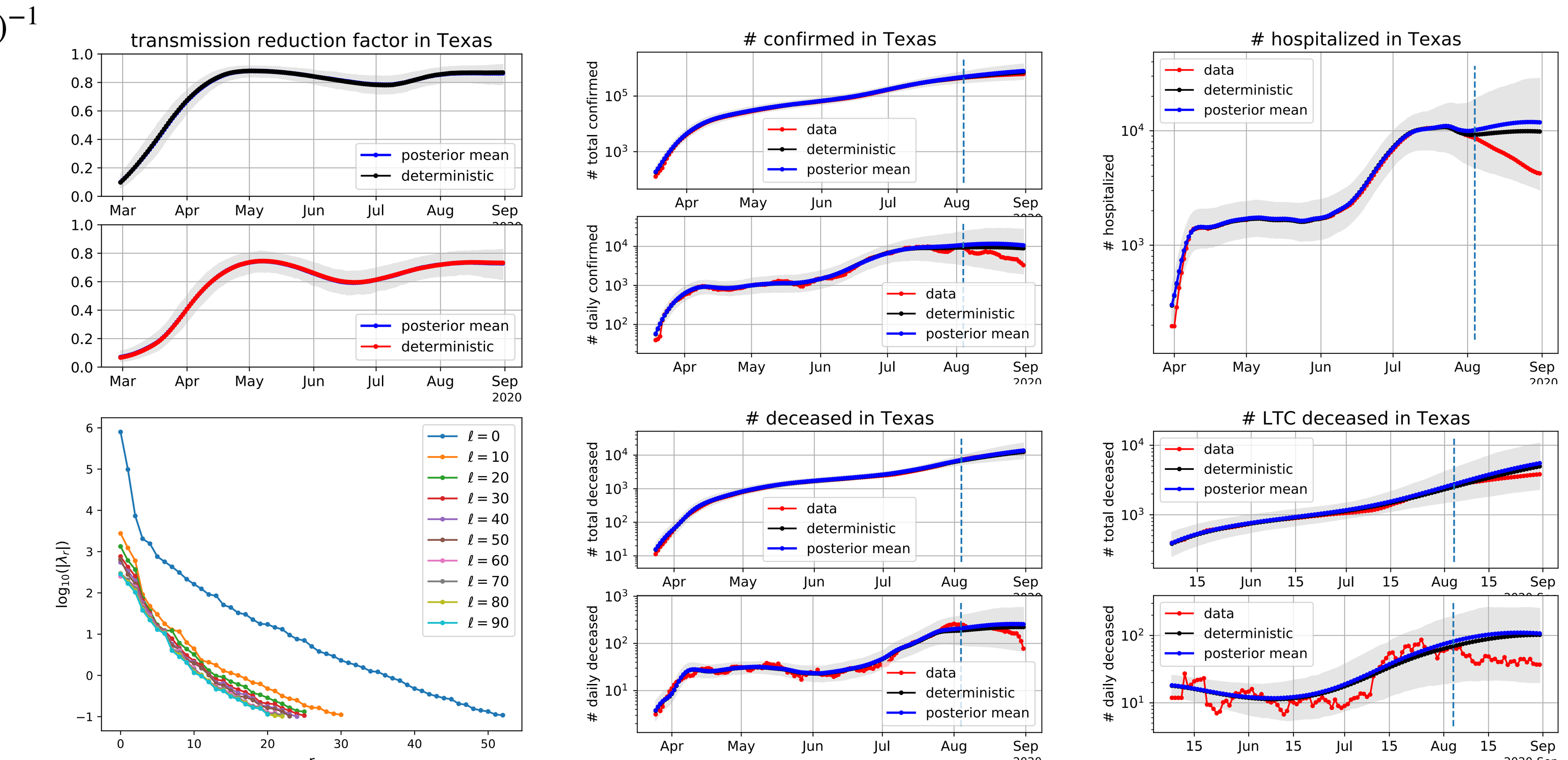
$$x \in \mathcal{N}(0, \mathscr{C}), \mathscr{C} = (-0.1\Delta + I)^{-2}$$
$$-\nabla \cdot (e^x \nabla u) = 0, \quad \text{in } (0,1)^2 .$$

ODE model with stochastic process

$$g_i(t) \in \mathcal{N}(g_i^*, \mathscr{C}_g), \mathscr{C}_g = (-s_g \Delta_t + s_I I)^{-1}$$
$$\alpha_i(t) = \frac{1}{2}\left(\tanh(g_i(t)) + 1\right) .$$



Experiment II: Inference of COVID-19



References

P. Chen, O. Ghattas. Projected Stein variational gradient descent. Advances in Neural Information Processing Systems, 2020.

P. Chen, K. Wu, O. Ghattas. Bayesian inference of heterogeneous epidemic models: Application to COVID-19 spread accounting for long-term care facilities. arXiv:2011.01058, 2020.

P. Chen, K. Wu, J. Chen, T. O'Leary-Roseberry, O Ghattas. Projected Stein variational Newton: A fast and scalable Bayesian inference method in high dimensions. Advances in Neural Information Processing Systems, 2019.

O. Zahm, T. Cui, K. Law, A. Spantini, Y. Marzouk. Certified dimension reduction in nonlinear Bayesian inverse problems, arXiv:1807.03712, 2018