高维情形下线性模型的泛化误差研究

寒假工作

Yukun Dong

目录

任务:

- 阅读 Learning with Reproducing Kernel Hilbert Spaces: Stochastic Gradient Descent and Laplacian Estimation，复现其中的一些图。
- 初读经验过程理论

- 延伸阅读：Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization.
- 延伸阅读：Alnur Ali, Edgar Dobriban, and Ryan J Tibshirani. The Implicit Regularization of Stochastic Gradient Flow for Least Squares.

RKHS

援引 Steinwart and Christmann (2014) 里面对 RKHS 的定义。

Definition 4.18. *Let $X \neq \emptyset$ and H be a $\mathbb{K}$-Hilbert function space over X, i.e., a $\mathbb{K}$-Hilbert space that consists of functions mapping from X into $\mathbb{K}$.*

*i) A function $k : X \times X \to \mathbb{K}$ is called a **reproducing kernel** of H if we have $k(\cdot, x) \in H$ for all $x \in X$ and the **reproducing property***

$$f(x) = \langle f, k(\cdot, x) \rangle$$

holds for all $f \in H$ and all $x \in X$.

*ii) The space H is called a **reproducing kernel Hilbert space (RKHS)** over X if for all $x \in X$ the Dirac functional $\delta_x : H \to \mathbb{K}$ defined by*

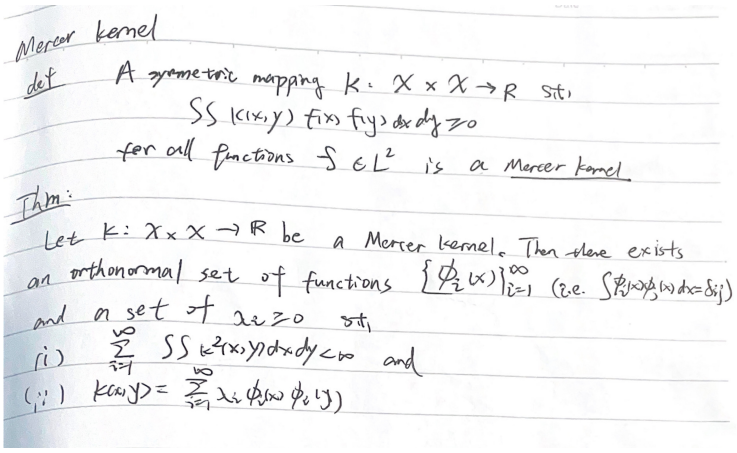
$$\delta_x(f) := f(x), \qquad f \in H,$$

is continuous.

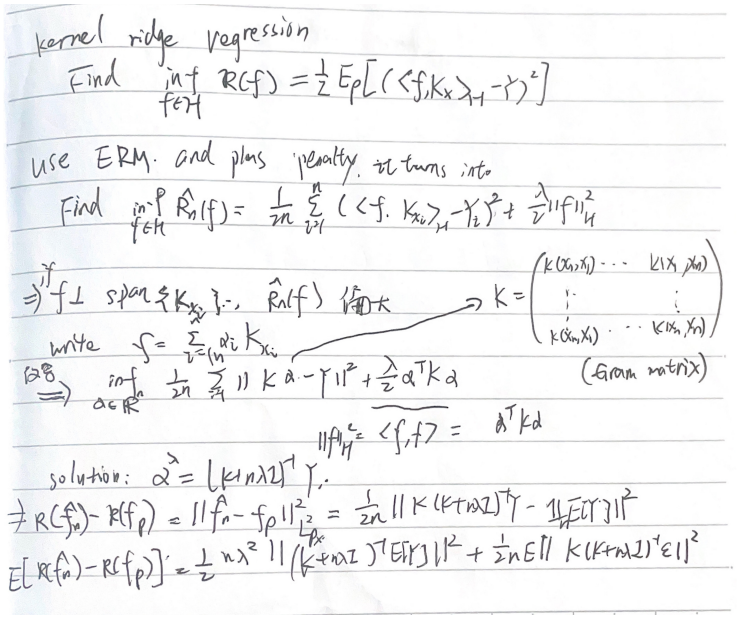
RKHS 有一些美好的性质。

- 最经典的，reproducing property
- 依范数收敛蕴含逐点收敛
- ...

Mercer kernel&Mercer Theorem

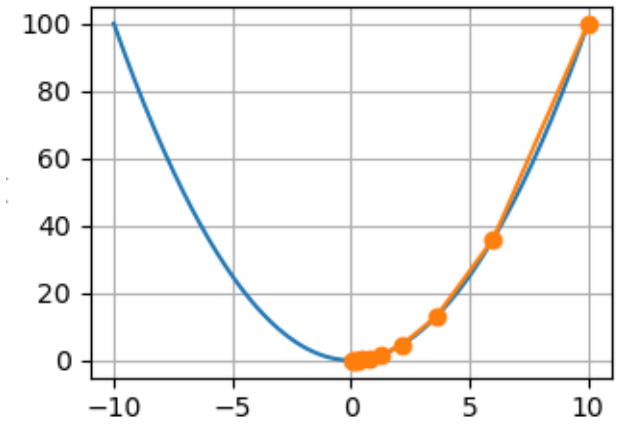


Kernel Ridge Regression

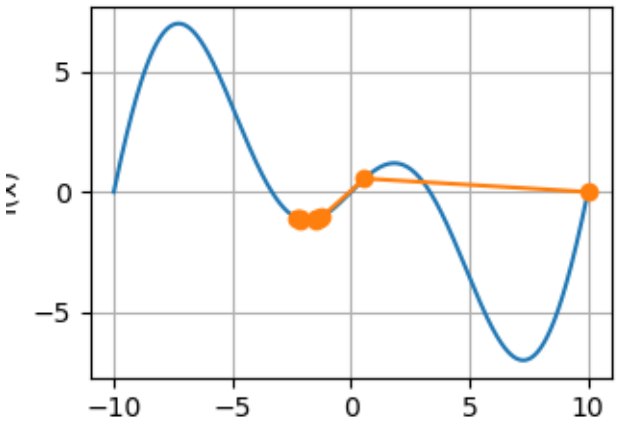


梯度下降 (Gradient Descent)

先用代码实现最简单的梯度下降。下图展现了目标函数 $y = x^2$，学习率 $\eta = 0.2$ 的梯度下降轨迹：



下图模拟了目标函数 $y = x * \cos(0.15\pi x)$，学习率 $\eta = 2$ 的梯度下降，可以发现无法收敛到全局最小：



下面是多元梯度下降。假设目标函数 $f : \mathbb{R}^d \to \mathbb{R}$ 的输入是一个 d 维向量 $x = [x_1, x_2, \ldots, x_d]^\top$。目标函数 $f(x)$ 有关 x 的梯度是一个由 d 个偏导数组

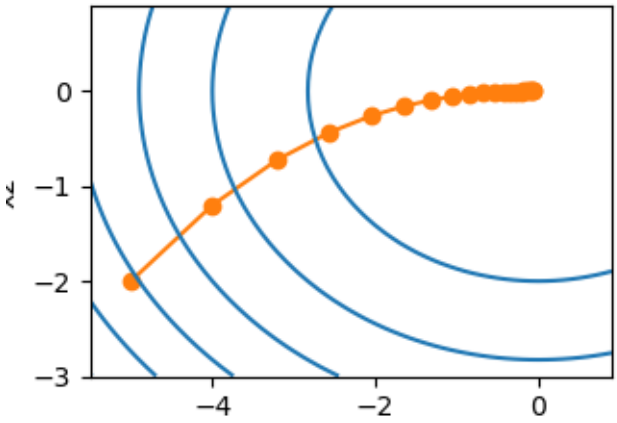
成的向量：

$$\nabla_x f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_d}\right]^\top.$$

通过梯度下降算法来不断降低目标函数 f 的值：

$$x \leftarrow x - \eta \nabla f(x).$$

构造一个输入为二维向量 $x = [x_1, x_2]^\top$ 和输出为标量的目标函数 $f(x) = x_1^2 + 2x_2^2$。那么，梯度 $\nabla f(x) = [2x_1, 4x_2]^\top$，观察梯度下降从初始位置 $[-5, -2]$ 开始对自变量 x 的迭代轨迹：



随机梯度下降 (Stochastic Gradient Descent)

目标函数定义为

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$

目标函数在 x 处的梯度计算为

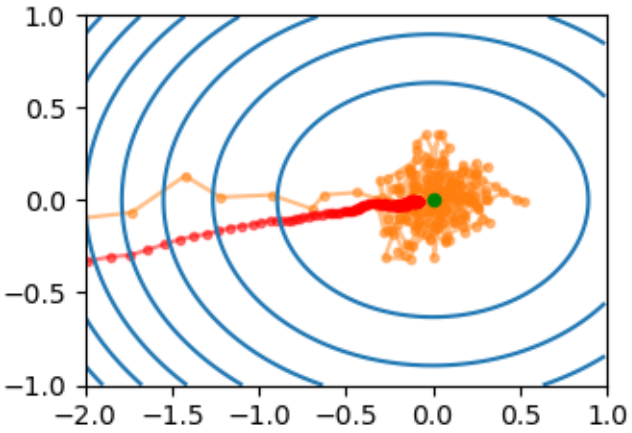
$$\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x).$$

在随机梯度下降的每次迭代中，我们随机均匀采样的一个样本索引 $i \in \{1, \ldots, n\}$，并计算梯度 $\nabla f_i(x)$ 来迭代 x：

$$x \leftarrow x - \eta \nabla f_i(x).$$

这样，每次迭代的计算开销从梯度下降的 $\mathcal{O}(n)$ 降到了常数 $\mathcal{O}(1)$。随机梯度 $\nabla f_i(x)$ 是对梯度 $\nabla f(x)$ 的无偏估计。

下面通过在梯度中添加均值为 0 的随机噪声来模拟随机梯度下降，以此来比较它与梯度下降的区别。这里 $f(x) = x_1^2 + 2 \times x_2^2$



可以看到，随机梯度下降中自变量的迭代轨迹相对于梯度下降总的来说更为曲折。上图中的红色轨迹代表做平均之后的参数曲线，即 $\overline{x}^{(i)} = \sum_{k=1}^{i} x^{(k)}$。在满足一些特定条件的 f 上，$\{\overline{x}^{(i)}\} \to x^*, \text{a.s.}$。不过要想 $\{\overline{x}^{(i)}\}$ 收敛到 x^*，往往需要较多的迭代次数。

MCMC&Langevin dynamics Sampling

贝叶斯方法通过使用 MCMC 来捕捉参数的不确定性，Langevin 动力学在更新参数时注入高斯噪声，这样做可以防止参数估计仅仅收敛到 MAP（最大后验众数）解决方案。取而代之，迭代会收敛到真实后验分布。由贝叶斯公式 $p(\theta|X) \propto p(\theta)\prod_{i=1}^{N} p(x_i|\theta)$，在每次迭代中选取 $n < N$ 个样本 $\{x_{t1}, \cdots, x_{tn}\}$，我们得到 SGD optimization 的迭代公式：

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)\right)$$

Langevin dynamics 的引入增加了噪声项，使得抽取样本的方差与后验分布的方差相同，亦即：

$$\Delta\theta_t = \frac{\epsilon}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)\right) + \eta_t,\ \eta_t \sim N(0, \varepsilon).$$

Xifara et al. (2014) 提出了 Metropolis-adjusted Langevin algorithm，相比没有校正的情况有更好地生成后验分布的能力。下图是在目标分布为高斯分布或香蕉形状时，在有无 Metropolis-adjustment 下利用 stochastic gradient Langevin dynamics 抽样的结果：

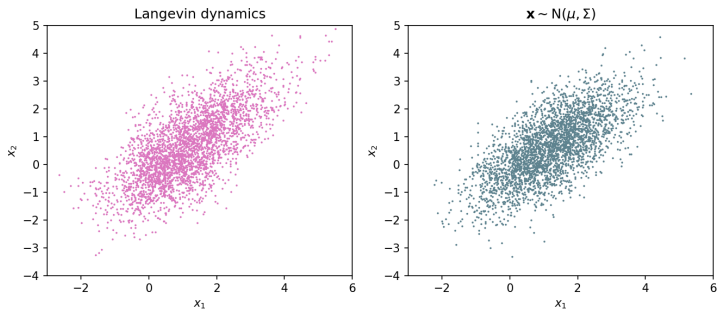


图 1: unadjusted Langevin dynamics: Gaussian

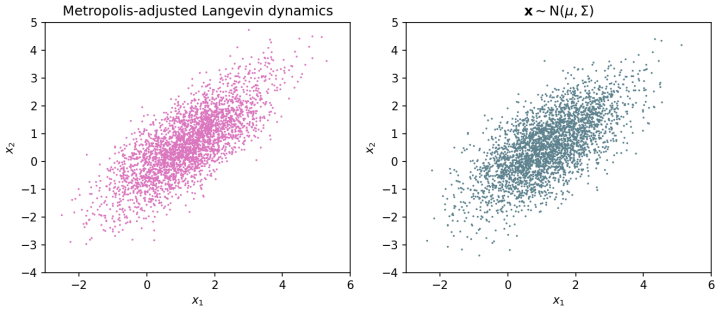


图 2: Metropolis-adjusted Langevin algorithm: Gaussian

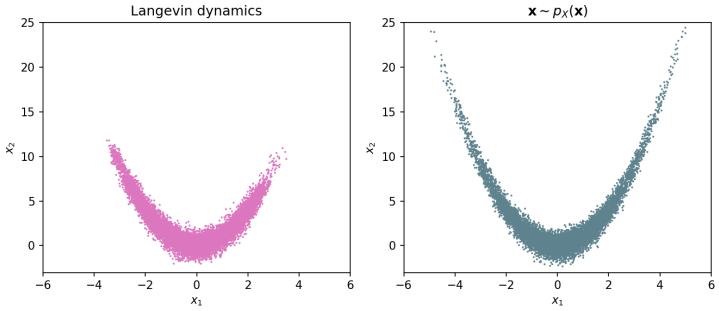


图 3: unadjusted Langevin dynamics: Banana

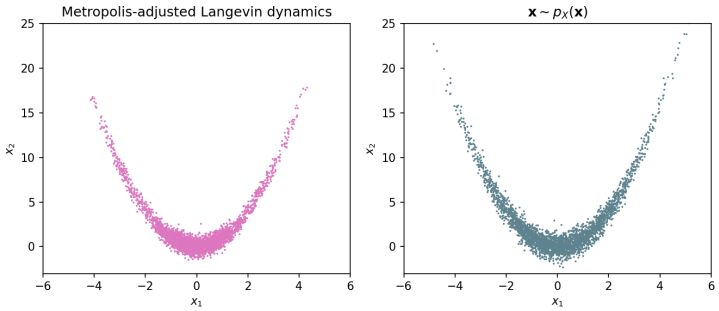


图 4: Metropolis-adjusted Langevin algorithm: Banana

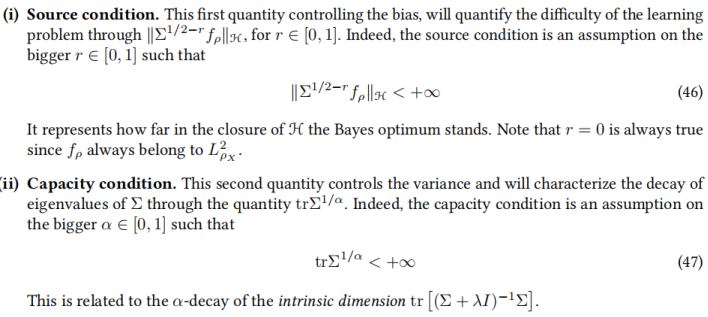
可见经过 Metropolis-adjustment 的算法有更好的抽样效果。代码部分参考了 Ali Siahkoohi 的仓库，我先前的 R 代码效果不佳，下一阶段将会致力于

弄清楚其原理。

待解决的问题

文章看下来自己的问题肯定很多，一些能具象化的问题罗列如下：

- P33，伊藤积分，SDE discreterization 有待了解；
- P44 的两个 condition 的意义是什么，这个在后面的文章里面也出现过；



(i) Source condition. This first quantity controlling the bias, will quantify the difficulty of the learning problem through $\|\Sigma^{1/2-r} f_\rho\|_{\mathcal{H}}$, for $r \in [0,1]$. Indeed, the source condition is an assumption on the bigger $r \in [0,1]$ such that

$$\|\Sigma^{1/2-r} f_\rho\|_{\mathcal{H}} < +\infty \tag{46}$$

It represents how far in the closure of $\mathcal{H}$ the Bayes optimum stands. Note that $r = 0$ is always true since f_ρ always belong to $L^2_{\rho_X}$.

(ii) Capacity condition. This second quantity controls the variance and will characterize the decay of eigenvalues of Σ through the quantity $\mathrm{tr}\Sigma^{1/\alpha}$. Indeed, the capacity condition is an assumption on the bigger $\alpha \in [0,1]$ such that

$$\mathrm{tr}\Sigma^{1/\alpha} < +\infty \tag{47}$$

This is related to the α-decay of the *intrinsic dimension* $\mathrm{tr}\left[(\Sigma + \lambda I)^{-1}\Sigma\right]$.

图 5: try

- P52 (ii) 是怎么推出来的？
- P136 infinitesimal generator of the Langevin diffusion 不懂是什么，上面那个式子涉及到 Markov 半群的微分，谱理论，根本不懂。

接下来的 to-do

- 弄清楚 Ali Siahkoohi 的代码原理
- 复现 P34 figure 5（Ali, Dobriban, and Tibshirani (2020) 的 figure 2）
- 第二篇文章要看懂需要补很多 literature，重读 if necessary
- 仔细阅读 Welling and Teh (2011) if necessary，这篇文章写的感觉很 intuitive，涵盖了 MCMC，贝叶斯，Langevin dynamics 较为易懂

参考文献

Ali, Alnur, Edgar Dobriban, and Ryan J. Tibshirani. 2020. “The Implicit Regularization of Stochastic Gradient Flow for Least Squares.” https://arxiv.org/abs/2003.07802.

Steinwart, I., and A. Christmann. 2014. *Support Vector Machines*. Information Science and Statistics. Springer New York. https://books.google.com/books?id=vFP4oQEACAAJ.

Welling, Max, and Yee Whye Teh. 2011. “Bayesian Learning via Stochastic Gradient Langevin Dynamics.” In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 681–88. ICML’11. Madison, WI, USA: Omnipress.

Xifara, T., C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami. 2014. “Langevin Diffusions and the Metropolis-Adjusted Langevin Algorithm.” *Statistics &Amp; Probability Letters* 91 (August): 14–19. https://doi.org/10.1016/j.spl.2014.04.002.