

高维情形下线性模型的泛化误差研究

2024 春季学期 week1,2 工作

Yukun Dong

目录

任务:	1
文章精读: Bayesian Learning via Stochastic Gradient Langevin	
Dynamics	2
SGD	2
Langevin Dynamics	2
Stochastic Gradient Langevin Dynamics	3
复现 Figure1	4
复现 Figure2.1	6
待解决的问题	7
接下来的 to-do	7
参考文献	8

任务:

- 精读并复现 Welling and Teh (2011), 是有关 bayes learning 和 langevin dynamics 的高引文章, 比较有启发性。

文章精读: *Bayesian Learning via Stochastic Gradient Langevin Dynamics*²

- 听报告: On Langevin Dynamics in Machine Learning - Michael I. Jordan。
- 弄清楚 Ali Siahkoobi 的代码原理
- (optional) 复现 P34 figure 5 (Ali, Dobriban, and Tibshirani (2020) 的 figure 2)

文章精读: Bayesian Learning via Stochastic Gradient Langevin Dynamics

SGD

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 $\{x_{t1}, \dots, x_{tn}\}$, 我们得到 SGD optimization 的迭代公式:

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

要保证序列能够收敛到似然函数的局部极大值处, 步长序列 $\{\epsilon_n\}$ 需要满足:

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \quad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty.$$

前者可以保证任何初值可以得到收敛结果, 后者使得轨迹不会在目标值附近一直跳跃。SGD 的问题在于它把握不了后验分布的随机性, 算法单单收敛到 MAP 或许是一种过拟合。

Langevin Dynamics

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:

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

文章精读: *Bayesian Learning via Stochastic Gradient Langevin Dynamics* 3

注意这里的步长是定值, 这是传统 MCMC 的要求, 以获得平稳的 markov chain。 ε 减小, 对应的 discretization error 减小, 拒绝率趋向于 0。

Stochastic Gradient Langevin Dynamics

结合 SGD 和 LD, updates 写为:

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

这个 update 既有 SGD, 又使得拒绝率收敛于 0. 在这种迭代下, θ_t 最终会达到后验分布。

需要注意的是有如下两点性质:

- θ_t 最终会达到后验分布
- 高斯噪声最终会主导 SGD 的噪声

对于这一点的证明见下面笔记:

Bayesian Learning via Stochastic Gradient Langevin Dynamics

SGLD $\begin{cases} \{\theta_t\} \text{ 趋向后验分布} \\ \eta_t \text{ (高斯噪声) 导 } h_t(\theta_t) \text{ (SGD的噪声)} \end{cases}$

maximize $\underset{\theta}{p(\theta)} \prod_{i=1}^N p(x_i | \theta)$
a-priori

$g(\theta) = \nabla \log p(\theta) + \frac{1}{N} \sum_{i=1}^N \nabla \log p(x_i | \theta)$: 真实的梯度的期望.

recall SGLD: $\Delta \theta_t = \frac{\eta_t}{\gamma} \left(\nabla \log p(\theta) + \frac{1}{N} \sum_{i=1}^N \nabla \log p(x_i | \theta_t) \right) + \eta_t$

其中 $\eta_t \sim N(0, \Sigma_t)$ $\sum \Sigma_t = \infty, \sum \Sigma_t^2 < \infty$.

$h_t(\theta) = \nabla \log p(\theta) + \frac{1}{N} \sum_{i=1}^N \nabla \log p(x_i | \theta) - g(\theta)$.

$h_t(\theta)$ 为零均值随机变量, 记为 $v(\theta)$

(下面要证明 $\{\theta_t\}$ 会趋向后验分布: 需证明一个序列 $\theta_{t_1}, \theta_{t_2}, \dots$ 会收敛到后验分布)

取定 $\varepsilon_0: 0 < \varepsilon_0 < 1, \exists t_1 < t_2 < \dots$ s.t. $\sum_{t=t_k+1}^{t_{k+1}} \Sigma_t \rightarrow \varepsilon_0$ as $k \rightarrow \infty$

则 $\|\sum_{t=t_k+1}^{t_{k+1}} \eta_t\|_2$ 为 $O(\sqrt{\varepsilon_0})$.

由于 $\varepsilon_0 < 1$, 可以选取 $\|\theta_{t_k} - \theta_{t_{k+1}}\|_2 < 1$, 对 $t_k \leq t \leq t_{k+1}$

考察 t_k 到 t_{k+1} 的 total stochastic gradient

$$\begin{aligned} \sum_{t=t_k+1}^{t_{k+1}} [g(\theta_t) + h_t(\theta_t)] \frac{\varepsilon_t}{\gamma} &= \frac{\varepsilon_0}{\gamma} g(\theta_{t_k}) + O(\varepsilon_0) + \underbrace{\sum_{t=t_k+1}^{t_{k+1}} \frac{\varepsilon_t}{\gamma} h_t(\theta_t)}_{O(\sqrt{\sum_{t=t_k+1}^{t_{k+1}} \varepsilon_t^2})} \\ &= \underbrace{\frac{\varepsilon_0}{\gamma} g(\theta_{t_k}) + O(\varepsilon_0)}_{\downarrow} < O(\sqrt{\varepsilon_0}). \end{aligned}$$

约等于 exact gradient step. 在 θ_{t_k} 处因噪声 η_t 而被 injected noise 干扰.

复现 Figure1

实验设定: 先验分布和数据生成的分布如下:

$$\theta_1 \sim N(0, \sigma_1^2); \quad \theta_2 \sim N(0, \sigma_2^2)$$

$$x_i \sim \frac{1}{2} N(\theta_1, \sigma_x^2) + \frac{1}{2} N(\theta_1 + \theta_2, \sigma_x^2)$$

其中 $\sigma_1^2 = 10, \sigma_2^2 = 1$ 和 $\sigma_x^2 = 2$ 。有 100 个数据点从模型中抽取, 其中 $\theta_1 = 0$ 和 $\theta_2 = 1$ 数据生成的参数, 而 $\theta_1 = 1, \theta_2 = -1$ 是一组能生成同样数据的参数, 参数之间有很强的负相关性。使用批量大小为 1, 并通过整个数据集进行了 100000 次迭代的 SGLD 算法。步长大小为 $\varepsilon_t = a(b+t)^{-\gamma}$, 其中 $\gamma = .55$ 并且 a 和 b 设置成使得 ε_t 在运行期间从 0.01 减少到 0.0001。

下面的图对比了实际的后验分布和 SGLD 的采样数据点:

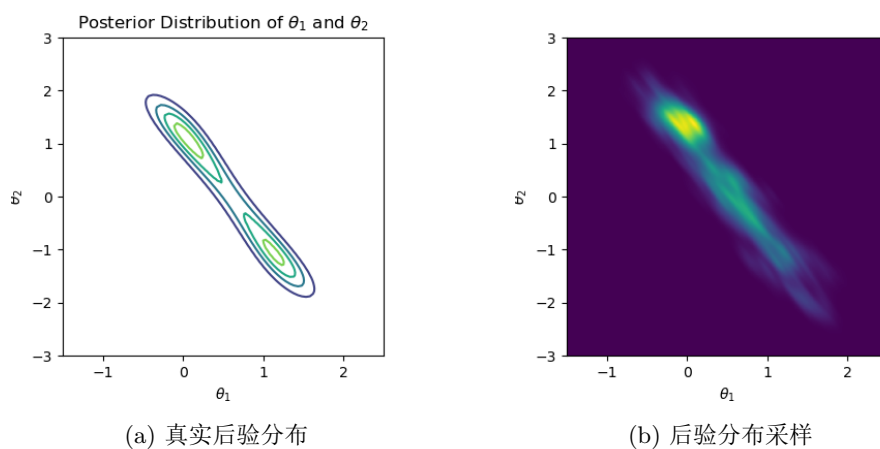


图 1: 我的模拟结果

与论文原图对比:

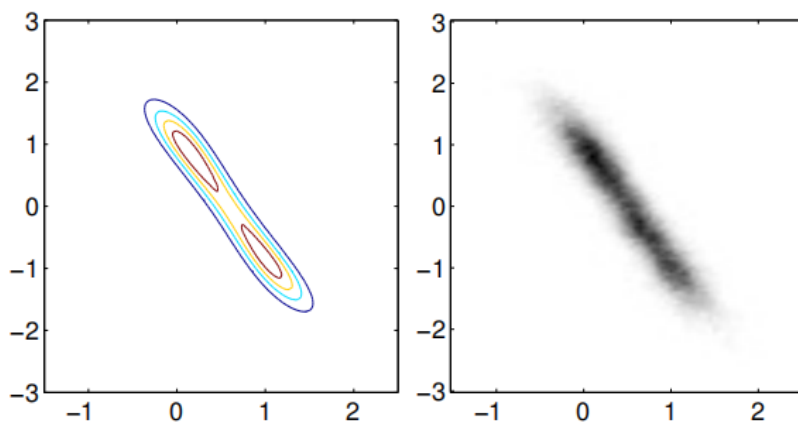


Figure 1. True and estimated posterior distribution.

图 2: 原文图 1

复现 Figure2.1

模拟原文图 2，比较了 injected noise 和 SGD 的 noise 随着迭代次数的变化。设定同上。图结果表明，在 $\{\varepsilon_t\}$ 满足 $\sum_{t=1}^{\infty} \varepsilon_t = \infty, \sum_{t=1}^{\infty} \varepsilon_t^2 < \infty$ 时，最终 injected noise 会主导 noise from SGD，这也是 $\{\theta_t\}$ 收敛到后验分布的保障。

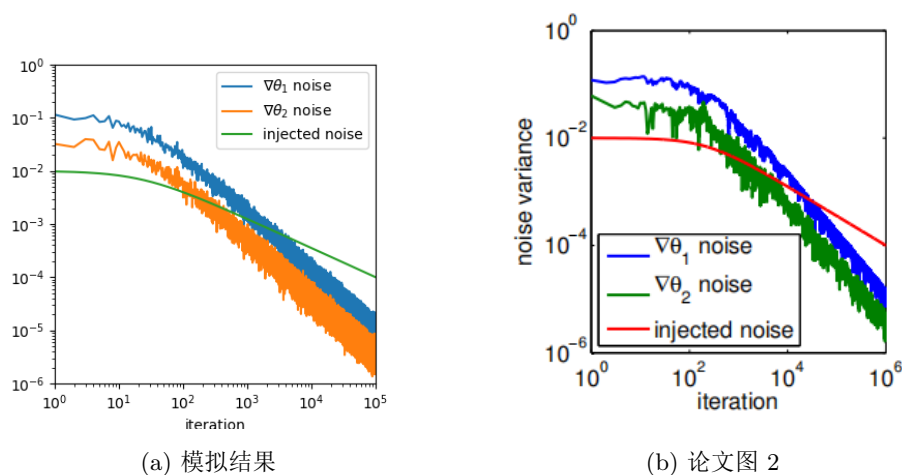


图 3: 我的模拟结果与原文结果对比

这个结果证实了在迭代次数变多的时候，SGD 的噪声在加性高斯噪声面前变得可忽略；此外，随着步长的减小，加上 MH 接受拒绝与否对于算法的影响微乎其微，因为最后拒绝率趋向于零。

待解决的问题

- Gradient flow 相关知识欠了解，有待读文章并写一些代码。
- 编程任务逐渐变难，以后要写出代码框架的话可能要学一下 pytorch。

接下来的 to-do

- 花大概 20 天左右仔细阅读 Bartlett, Montanari, and Rakhlin (2021) Deep learning: a statistical viewpoint, 讲的是深度学习的统计思想，包含隐正则化，min-norm least squares, gradient flow 以及 RKHS, 值得细品。有待阅读的章节：1 Introduction, 2 Generalization and uniform convergence, 3 Implicit regularization, 4 Benign overfitting, 6 Generalization in the linear regime。
- 了解 Gradient flow 相关知识，争取复现 Ali, Dobriban, and Tibshirani (2020) 的 figure 2。

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>.
- Bartlett, Peter L., Andrea Montanari, and Alexander Rakhlin. 2021. “Deep Learning: A Statistical Viewpoint.” <https://arxiv.org/abs/2103.09177>.
- 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.