Stochastic Gradient hamiltonian Monte Carlo

竹田航太

2021 年 3 月 30 日

概要

機械学習や最適化の分野で非凸最適化の需要から生まれたのが Stochastic Gradient Method(確率的勾配法) である．その技術を HMC に取り入れたのが Stochastic Gradient HMC である．この手法では従来の HMC に比べポテンシャルに対する凸性の制約が散逸性に緩和される．また，SDE の数値計算とも関係が非常に深く Markov 過程の理論を使って定常分布への収束を示す．ここでは，HMC に関する基礎知識は前提とする．

1 基本のアイデア

サンプリング問題において目的分布のポテンシャルが以下のように和の形で表される場合を考える．$N \in \mathbb{N}$ に対して，

$$U(x) = \sum_{i=1}^{N} U_i(x) \qquad\qquad x \in \mathbb{R}^d \tag{1.1}$$

$\pi_U(x) \propto e^{-U(x)}$ からのサンプリングを考える．運動量 p を加えて相空間上の分布に拡張する $\pi_H(x,p) \propto \exp(-U(x) - p^\top M^{-1} p)$, M は対称正定値.

1.1 Stochastic Gradient

N が大きい場合，$U(x)$ の勾配を計算するコストが高くなるので代わりに少数の部分ポテンシャルで近似する．$2^{\{1,\cdots,N\}}$ 上の確率測度 ν に従って，index 集合の部分集合 $\mathcal{D} \subset \{1,\cdots,N\}$ をランダムにサンプルする．部分 index 集合 $\mathcal{D}$ から，以下のように $\nabla U(x)$ を近似する．

$$\nabla \tilde{U}_{\mathcal{D}}(x) = \frac{N}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \nabla U_i(x) \tag{1.2}$$

上記の勾配近似に次を仮定する．

Assumption 1.1. $\exists V : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ *s.t.*

$$\nabla \tilde{U}_{\mathcal{D}}(x) \approx \nabla U(x) + \mathcal{N}(0, V(x))$$

ただし，$\mathcal{N}(m, s)$ は平均 m, 分散 $s > 0$ の正規分布.

次の過程は数学的により厳密ではある．ここでは扱わず，後に紹介する.

Assumption 1.2 (厳密な仮定)**.** 以下の *2* つを満たすとする.

(1) $\mathbb{E}_\nu[\nabla \tilde{U}_{\mathcal{D}}(x)] = U(x)$; 不偏推定
(2) $\exists C > 0$ *s.t.* $\mathbb{E}_\nu[\|\nabla \tilde{U}_{\mathcal{D}}(x) - U(x)\|^2] \leq C$

1.2 Naïve SGHMC

基本的な HMC の Alg は 3step で構成される．運動量リサンプリング，Hamilton flow, MH-check である．このうちの Hamilton flow の従う方程式は以下

$$\begin{cases} dx_t &= M^{-1} p_t dt \\ dp_t &= -\nabla U(x_t) dt \end{cases} \tag{1.3}$$

まず次のような ∇U を $\nabla \tilde{U}$ で置き換えた単純な SGHMC が思いつく.

$$\begin{cases} dx_t &= M^{-1} p_t dt \\ dp_t &= -\nabla \tilde{U}(x_t) dt = -\nabla U(x) + \sqrt{2B} dW_t \end{cases} \tag{1.4}$$

ただし，W_t は $\mathbb{R}^d$ 上のブラウン運動．$B(x) = \frac{1}{2}\epsilon V(x)$

Remark 1.3. *Naïve SGHMC* の設定は例えばランダムな風が吹いている中での調和振動に対応する.

Theorem 1.4 (Naïve SGHMC の不完全性)**.** *(1.4)* の *SDE* に対応する *Fokker-Plank* 方程式の解を $p_t(x, p)$ と書く．また，$\mathbb{R}^{2d}$ 上の確率分布 ρ に対してエントロピー h を次で定義する.

$$H(\rho) = -\int_{\mathbb{R}^{2d}} f(d\rho)$$

ただし, $f(t) = t \log t$．また, $p_t(x, p)$ に次を仮定する．$|p_t(x,p)|, |\nabla p_t(x,p)|, |\nabla p_t(x,p)||\log p_t(x,p)| \to 0$ *if* $(x,p) \to \infty$

このとき，エントロピーは次のように増大する.

$$\partial_t p_t(x,p) = \int f''(p_t)(\nabla_p p_t(x,p))^\top B(x)(\nabla_p p_t(x,p)) dx dp$$
$$\geq 0$$

1.3 SGHMC with Friction

Naïve SGHMC ではノイズの影響で分布がどんどん拡散していくので friction(摩擦) 項を加えてコントロールする．

$$
\begin{cases} dx_t &= M^{-1}p_t dt \\ dp_t &= -\nabla U(x) - BM^{-1}p_t + \sqrt{2B}dW_t \end{cases} \tag{1.5}
$$

Remark 1.5. この *friction* 項はランダムに吹く風による抵抗を導入していることに対応する．

Theorem 1.6. $\pi_H(x,p) \propto \exp(-H(x,p))$ は *(1.5)* の唯一の定常分布．

Proof. $\pi_H(x,p)$ が (1.5) に対応する Fokker-Plank 方程式の定常解となることを示す．まず，以下のように G, D を定める．

$$
G = \begin{bmatrix} O & -I \\ I & O \end{bmatrix}, D = \begin{bmatrix} O & O \\ O & B \end{bmatrix}
$$

すると (1.5) は次のようにかける．

$$
\begin{aligned}
d\begin{bmatrix} x \\ p \end{bmatrix} &= -\begin{bmatrix} O & -I \\ I & B \end{bmatrix}\begin{bmatrix} \nabla U(x) \\ M^{-1}p \end{bmatrix} + \sqrt{2D}dW_t \\
&= -[D+G]\nabla H(x,p)dt + \sqrt{2D}dW_t
\end{aligned}
$$

これに対応する Fokker-Plank 方程式は

$$
\begin{aligned}
\partial_t p_t(x,p) &= \nabla([D+G]\nabla H(x,p)p_t(x,p)) + \nabla \cdot \nabla D p_t(x,p) \\
&= \nabla([D+G][\nabla H(x,p)p_t(x,p)) + \nabla p_t(x,p)])
\end{aligned}
$$

最後の変形で $\nabla \cdot G\nabla = 0$ を使った．

$p_t(x,p) = \pi_H(x,p)$ は $\nabla H(x,p)p_t(x,p)) + \nabla p_t(x,p) = 0$ を満たすので $\partial_t p_t(x,p) = 0$ となり，Fokker-Plank 方程式の定常分布である． □

参考文献

[1] Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient hamiltonian monte carlo for non-convex stochastic optimization: Non-asymptotic performance bounds and momentum-based acceleration, 2020.

[2] Tianqi Chen, Emily B. Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo, 2014.