通信理論に特化した深層学習
第5回ゼミ資料
確率的局所最小化法

豊橋技術科学大学
電気・電子情報工学系
准教授　竹内啓悟

学習とは何か？

学習の目標

学習すべきパラメータを$\boldsymbol{\theta} \in \mathbb{R}^n$、評価用データを表す確率変数を$Z$、損失関数を$f(Z;\boldsymbol{\theta})$とする。期待損失$\mathbb{E}_Z[f(Z;\boldsymbol{\theta})]$を最小にするという意味で、未知の入力に対する出力が目標に最も近づくように、訓練用データを使ってパラメータ$\boldsymbol{\theta} \in \mathbb{R}^n$を最適化する。

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \mathbb{E}_Z[f(Z;\boldsymbol{\theta})]$$

学習の方法

大きさDの訓練データ$\{z_d\}_{d=1}^D$に基づく損失関数の経験平均が最小になるように、パラメータを最適化する。

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \frac{1}{D}\sum_{d=1}^{D} f(z_d;\boldsymbol{\theta})$$

総和で表現される多変数関数の最小化問題

大域的最適解を見つけるのは困難なので、局所最適解を探す。

学習とは、総和で表現される大自由度関数の局所最適化である。

局所最適化と過学習への対策

過学習は、一般に学習すべきパラメータ数に対して訓練用データが不足するときに発生する。

層数20、層当たりのユニット数500の全結合型ネットワーク

$$\text{重みの総数} = 20 \times 500^2 = 5 \times 10^6$$

ビッグデータとは名ばかりで、実際にはスモールデータとみなした方がよい。

過学習への対策

- 問題の構造を利用して、ネットワークの構造を制約する。

  深層学習が、画像、音声、自然言語処理で成功した理由

- 良い局所最適解に到達する初期値を与える。

  メッセージ伝播法を参考にする。

深層学習は万能ではなく、過学習を回避する人間の知恵が必要

バッチ学習

訓練データを全てまとめて使用する学習方法

目的関数

$$f(\boldsymbol{\theta}) = \sum_{d=1}^{D} f(z_d; \boldsymbol{\theta}).$$

勾配ベクトル

$$\nabla f = \left(\frac{\partial f}{\partial \theta_1}, \cdots, \frac{\partial f}{\partial \theta_n} \right)^{\mathrm{T}}.$$

勾配方向に微小移動すると、関数値は最大化される。

∵ 微小な$\epsilon > 0$と任意の単位ベクトル$\boldsymbol{u} \in \mathbb{R}^n$に対して、

$$f(\boldsymbol{\theta} + \epsilon \boldsymbol{u}) = f(\boldsymbol{\theta}) + \epsilon \nabla f(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{u} + \mathcal{O}(\epsilon^2) \leq f(\boldsymbol{\theta}) + \epsilon \|\nabla f(\boldsymbol{\theta})\| + \mathcal{O}(\epsilon^2)$$

コーシー・シュワルツの不等式から、等号成立は$\boldsymbol{u} = \nabla f(\boldsymbol{\theta}) / \|\nabla f(\boldsymbol{\theta})\|$に限る。

勾配降下法

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \alpha \nabla f(\boldsymbol{\theta}^{t-1}).$$

適切な初期値$\boldsymbol{\theta}^0 \in \mathbb{R}^n$と学習率（Learning rate）$\alpha > 0$を設定すると、
$\boldsymbol{\theta}^t$は関数fの局所最小値を与える解に収束する。

ミニバッチ学習

学習時に使用する計算資源の並列数等にしたがって、全訓練データを
ミニバッチと呼ばれる小さなサイズ$\widetilde{D}$のデータ集合に分割する。

$$\{1,\dots,D\} = \cup_{i=1}^{D/\widetilde{D}} \mathcal{D}_i, \qquad |\mathcal{D}_i| = \widetilde{D}, \qquad \mathcal{D}_i \cap \mathcal{D}_j = \emptyset \text{ for } i \neq j.$$

確率的勾配降下（Stochastic gradient descent, SGD）法

パラメータの更新の度に、ミニバッチiをランダムに選びなおす。

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \alpha \nabla f_i(\boldsymbol{\theta}^{t-1}), \qquad f_i(\boldsymbol{\theta}) = \sum_{d \in \mathcal{D}_i} f(z_d; \boldsymbol{\theta}).$$

パラメータ更新

ミニバッチの数$D/\widetilde{D}$にパラメータ更新回数が制約されないように、
ミニバッチの再利用を許す。

エポック数：ミニバッチ当たりの再利用回数

訓練用データ数10000、ミニバッチサイズ$\widetilde{D} = 50$、エポック数3の場合、
パラメータの更新は$3 \times 10000/50 = 600$回行われる。

AdaGrad

パラメータ$\boldsymbol{\theta}$の要素jごとに学習率を適用的に制御して収束を早める。

$$\theta_j^t = \theta_j^{t-1} - \frac{\alpha}{\sqrt{v_j^t}} \frac{\partial f_i}{\partial \theta_j}(\boldsymbol{\theta}^{t-1}),$$

$$v_j^t = v_j^{t-1} + \left(\frac{\partial f_i}{\partial \theta_j}(\boldsymbol{\theta}^{t-1}) \right)^2, \qquad v_j^0 = \epsilon > 0.$$

解釈

過去のパラメータの変化量が多いと、学習率は適用的に小さくなる。

過去の変化量が少ないパラメータほど、優先して更新すべきという直観に基づく。

[5-1] J. Duchi, E. Hazan, and Y. Singer, “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization,” *J. Mach. Learn. Res.,* vol. 12, pp. 2121-2159, Jul. 2011.

Adam

AdaGradにおける勾配と学習率の制御パラメータv_j^tをそれぞれ
勾配の平均と二次モーメントとの移動平均不偏推定量に取る。

勾配の平均の推定量　$\beta_1 \in [0,1)$は平均に関する忘却係数

$$\hat{m}_j^t = \frac{m_j^t}{1-\beta_1^t}, \qquad m_j^t = \beta_1 m_j^{t-1} + (1-\beta_1)\frac{\partial f_i}{\partial \theta_j}(\boldsymbol{\theta}^{t-1}), \qquad m_j^0 = 0.$$

勾配の二次モーメントの推定量

$$\hat{v}_j^t = \frac{v_j^t}{1-\beta_2^t}, \qquad v_j^t = \beta_2 v_j^{t-1} + (1-\beta_2)\left(\frac{\partial f_i}{\partial \theta_j}(\boldsymbol{\theta}^{t-1})\right)^2, \qquad v_j^0 = 0.$$

$\beta_2 \in [0,1)$は二次モーメントに関する忘却係数

パラメータ更新式

$$\theta_j^t = \theta_j^{t-1} - \frac{\alpha}{\sqrt{\hat{v}_j^t + \epsilon}}\hat{m}_j^t.$$

$\alpha = 0.001,\ \beta_1 = 0.9,\ \beta_2 = 0.999,\ \epsilon = 10^{-8}$が推奨されている。

[5-2] D. P. Kingma and J. L. Ba, “Adam: A Method for Stochastic Optimization,” in *Proc. 3rd Int. Conf. Learn. Rep.*, San Diego, CA, USA, May 2015.

不偏推定量であることの確認

反復tにおいて使用されたミニバッチを$\mathcal{D}_{i_t}$とする。

m_j^tの定義式から、

$$m_j^t = (1-\beta_1)\sum_{t'=1}^{t}\beta_1^{t-t'}\frac{\partial f_{i_{t'}}}{\partial \theta_j}\left(\boldsymbol{\theta}^{t'-1}\right).$$

両辺の期待値を取ると、

$$\mathbb{E}\left[m_j^t\right] = (1-\beta_1)\sum_{t'=1}^{t}\beta_1^{t-t'}\mathbb{E}\left[\frac{\partial f_{i_{t'}}}{\partial \theta_j}\left(\boldsymbol{\theta}^{t'-1}\right)\right]$$

$$\approx \mathbb{E}\left[\frac{\partial f_i}{\partial \theta_j}(\boldsymbol{\theta}^{t-1})\right](1-\beta_1)\sum_{t'=1}^{t}\beta_1^{t-t'} = (1-\beta_1^t)\mathbb{E}\left[\frac{\partial f_i}{\partial \theta_j}(\boldsymbol{\theta}^{t-1})\right].$$

近似の導出では、訓練用データの同一分布性と反復当たりのパラメータの変化量が微小であることとを仮定した。それゆえ、

$$\mathbb{E}\left[\hat{m}_j^t\right] \approx \mathbb{E}\left[\frac{\partial f_i}{\partial \theta_j}(\boldsymbol{\theta}^{t-1})\right]. \quad \text{同様に、} \mathbb{E}\left[\hat{v}_j^t\right] \approx \mathbb{E}\left[\left(\frac{\partial f_i}{\partial \theta_j}(\boldsymbol{\theta}^{t-1})\right)^2\right].$$

Adamは勾配ベクトルの各要素の大きさを1に近づけることを狙っている。

問題例

最小化問題

$$\min_{x,y\in\mathbb{R}} f(x,y)\,.$$

目的関数

$$f(x,y)=\frac{1}{100}\sum_{d=1}^{100} f(x,y,a_d)\,,\qquad f(x,y,a)=ax^2+y^2.$$

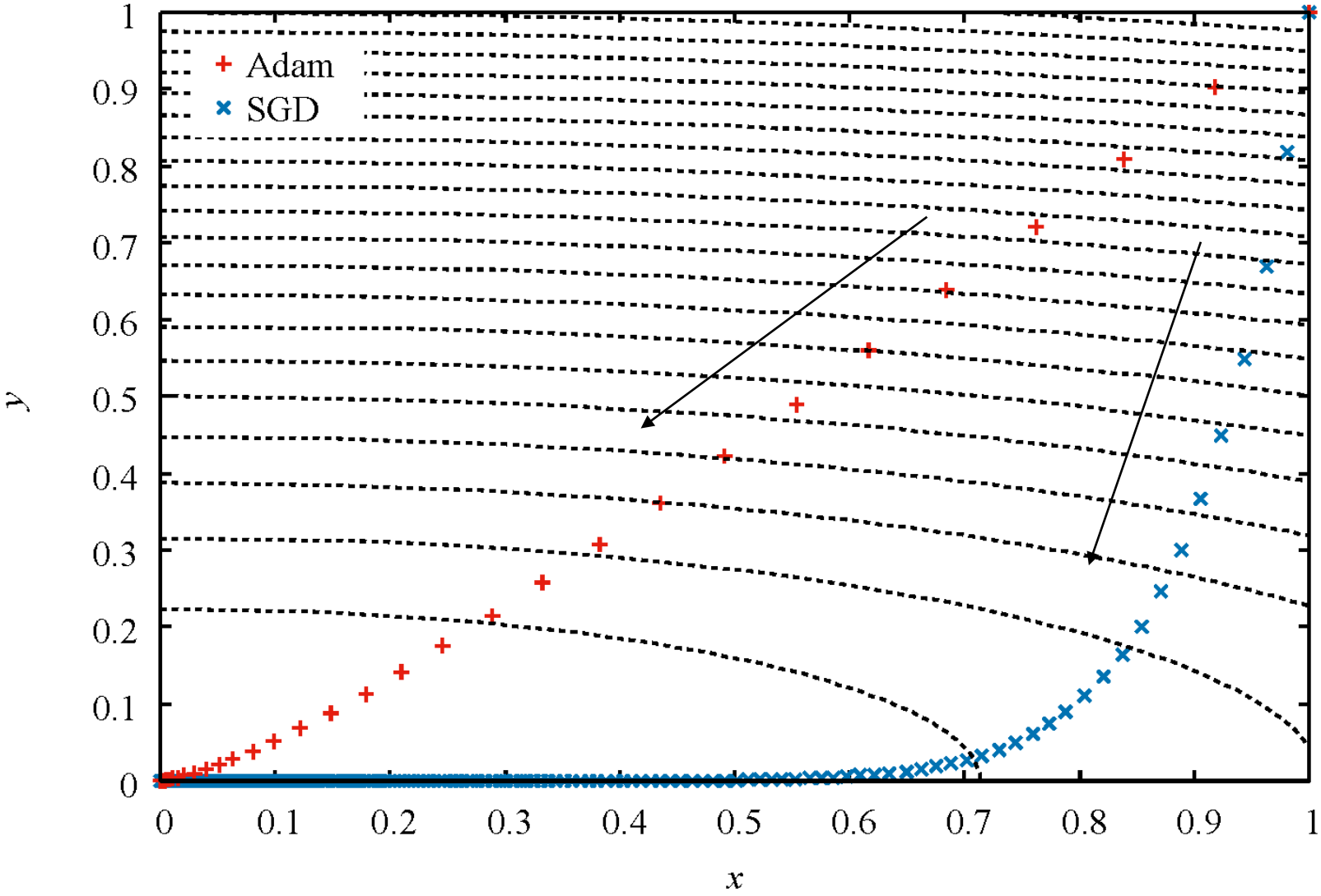
a_d : 区間$[0,0.2)$上の一様乱数$(\mathbb{E}[a_d]=0.1)$

目的関数の特徴

y軸方向の勾配が、x軸方向の勾配に比べて大きい。

y軸方向の勾配に合わせて学習率を設定すると、
x軸方向の変化が小さくなりすぎる。

シミュレーション（反復100回ごとのプロット）



初期値$(1, 1)$、学習率$\alpha = 0.001$、ミニバッチサイズ$\tilde{D} = 1$