Statistical analysis of stochastic gradient methods for generalized linear models

Panos Toulis
ptoulis@fas.harvard.edu
Harvard University,
Department of Statistics

Joint work with Jason Rennie (Google) & Edoardo M Airoldi (Harvard)

INTRODUCTION

FOCUS OF THIS WORK

- Stochastic Gradient Descent (SGD): computationally attractive, slow convergence, great empirical performance.
- As a *statistical* estimation method, still not well-understood.
- Say $\boldsymbol{\theta}_n^{\text{sgd}}$ is the output of SGD given observed data $\boldsymbol{y}$ generated by a model with true parameters $\boldsymbol{\theta}^*$; we ask:
  - What is the bias $\mathbb{E}\left(\boldsymbol{\theta}_n^{\text{sgd}} - \boldsymbol{\theta}^*\right)$?
  - What is the variance $\text{Var}\left(\boldsymbol{\theta}_n^{\text{sgd}}\right)$?
  - How to optimally set the learning rate?
- We also want to consider SGD with *explicit* and *implicit* updates and provide a meaningful comparison.

PROBLEM AND NOTATION

GLM FAMILY

At every time step indexed by n, assume the following data generating process:

- $\boldsymbol{x}_n \sim G$ sampled iid, $\in \mathbb{R}^p$ (features)
- $y_n \sim f(y_n; \boldsymbol{x}_n^\intercal \boldsymbol{\theta}^*, \psi) \in \mathbb{R}$ (outcome)

  such that $f(\cdot)$ is a density in the *exponential family* and

$$\mathbb{E}\left(y_n | \boldsymbol{x}_n\right) = h(\boldsymbol{x}_n^\intercal \boldsymbol{\theta}^*) \tag{1}$$

  where $h(\cdot)$ is the (monotone) *link* function, $\boldsymbol{\theta}^* \in \mathbb{R}^p$ are the unknown model parameters, and $\psi > 0$ is the dispersion parameter.

Our **goal** is to *estimate* $\boldsymbol{\theta}^*$ given observations $(y_i, \boldsymbol{x}_i)$, indexed by $i = 1, \cdots N$.

KNOWN PROPERTIES OF EXPONENTIAL FAMILY/GLMS.

Let $\ell(\boldsymbol{\theta}; y_n, \boldsymbol{x}_n)$ be the log-likelihood of $\boldsymbol{\theta}$ for observation $(y_n, \boldsymbol{x}_n)$. The following hold for a GLM:

$$\nabla \ell(\boldsymbol{\theta}; y_n, \boldsymbol{x}_n) = \frac{1}{\psi}\left(y_n - h(\boldsymbol{x}_n^{\intercal}\boldsymbol{\theta})\right)\boldsymbol{x}_n \tag{2}$$

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -\mathbb{E}\left(\nabla\nabla \ell(\boldsymbol{\theta}; y_n, \boldsymbol{x}_n)\right) = \frac{1}{\psi}\mathbb{E}\left(h'(\boldsymbol{x}_n^{\intercal}\boldsymbol{\theta})\boldsymbol{x}_n\boldsymbol{x}_n^{\intercal}\right) \tag{3}$$

- Equation (2) gives the gradient of the log-likelihood in terms of **“observed - expected”** of the sufficient statistics.
- Equation (3) gives the Fisher information matrix. The value $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)^{-1}$ is the theoretically **best** possible variance we can achieve if we try to (unbiasedly) estimate $\boldsymbol{\theta}^*$.

ITERATIVE ESTIMATION OF GLMS

SGD PROCEDURES

The *explicit* SGD updates are given by

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n \left(y_n - h(\boldsymbol{x}_n^\intercal \boldsymbol{\theta}_{n-1})\right) \boldsymbol{x}_n \tag{4}$$

The *implicit* SGD updates are given by

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n \left(y_n - h(\boldsymbol{x}_n^\intercal \boldsymbol{\theta}_{n})\right) \boldsymbol{x}_n \tag{5}$$

Remark #1. After n iterations, both procedures provide an *estimate* of $\boldsymbol{\theta}^*$:

- $\boldsymbol{\theta}_n^{\mathrm{sgd}}$ of the explicit updates is the *explicit SGD estimator* of $\boldsymbol{\theta}^*$.
- Similary, $\boldsymbol{\theta}_n^{\mathrm{im}}$ is the *implicit SGD estimator* of $\boldsymbol{\theta}^*$.

ITERATIVE ESTIMATION OF GLMS

SGD PROCEDURES

The *explicit* SGD updates are given by

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n \left(y_n - h(\boldsymbol{x}_n^\intercal \boldsymbol{\theta}_{n-1}) \right) \boldsymbol{x}_n \tag{6}$$

The *implicit* SGD updates are given by

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n \left(y_n - h(\boldsymbol{x}_n^\intercal \boldsymbol{\theta}_n) \right) \boldsymbol{x}_n \tag{7}$$

Remark #2. Implicit methods are less well-studied.

- Similar ideas have been used in numerical analysis e.g., Crank-Nicolson method (1947), to solve PDE.
- The NLMS algorithm in signal processing (Nagumo & Noda, 1967) uses an implicit method for linear regression.
- Recent interest due to stability of the method (Kivinen, 1996), (Kivinen et al., 2006), (Kulis & Bartlett, 2010).

EXAMPLE 1 : NORMAL MODEL

Assume

$$\boldsymbol{x}_n \sim G \text{ and } \boldsymbol{X}_n = \boldsymbol{x}_n \boldsymbol{x}_n^\intercal \text{ (possibly random)}$$
$$y_n \sim \mathcal{N}(\boldsymbol{x}_n^\intercal \boldsymbol{\theta}^*, \sigma^2)$$

The explicit SGD update is,

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n \underbrace{(y_n - \boldsymbol{x}_n^\intercal \boldsymbol{\theta}_{n-1})}_{\text{observed - expected}} \boldsymbol{x}_n = (\boldsymbol{I} - a_n \boldsymbol{X}_n)\boldsymbol{\theta}_{n-1} + a_n y_n \boldsymbol{x}_n$$

The implicit SGD can be derived analytically as,

$$\boldsymbol{\theta}_n = (\boldsymbol{I} + a_n \boldsymbol{X}_n)^{-1}(\boldsymbol{\theta}_{n-1} + a_n y_n \boldsymbol{x}_n)$$

The latter update is known as the “Normalized Least Mean Squares” filter.

Example 2: Poisson Regression

Assume

$$y_n \sim \mathrm{Pois}(e^{\boldsymbol{x}_n^\intercal \boldsymbol{\theta}^*})$$

The explicit SGD is,

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n \underbrace{(y_n - e^{\boldsymbol{x}_n^\intercal \boldsymbol{\theta}_{n-1}})}_{\text{observed - expected}} \boldsymbol{x}_n$$

The implicit SGD is,

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n (y_n - e^{\boldsymbol{x}_n^\intercal \boldsymbol{\theta}_n}) \boldsymbol{x}_n$$

- Explicit SGD is problematic in this model because of the non-linearity of the score function.
- The implicit update cannot be derived analytically. However, we will show how it can be *computed efficiently*.

EXAMPLE 2: POISSON REGRESSION (CONTINUED)

INSTABILITY OF **explicit** SGD

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n(y_n - e^{\boldsymbol{x}_n^\intercal \boldsymbol{\theta}_{n-1}})\boldsymbol{x}_n$$

- Assume one-dimensional case, for which $\theta_0 = 0, x_0 = x_1 = a_0 = 1, y_1 = 1001$, then the update is:

$$\theta_1 = 0 + 1(1001 - 1)1 = 1000$$

- In the next iteration assume that $y_2 = 500, a_1 = 0.5$, then:

$$\theta_2 = 1000 + 0.5(500 - e^{1000})1 = -\infty$$

The problem is that the starting point ($\theta_0 = 0$) was far away from the observation $y_1 = 1001$ and the learning rate was not small enough to prevent a large update (misspecification).

EXAMPLE 2: POISSON REGRESSION (CONTINUED)

STABILITY OF **implicit** SGD

Implicit update for Poisson regression model

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + a_n(y_n - e^{\boldsymbol{x}_n^\intercal \boldsymbol{\theta}_n})\boldsymbol{x}_n$$

- The implicit update would solve:

$$\theta_1 = 0 + 1(1001 - \underbrace{e^{\theta_1}}_{\text{implicit}})1$$

and so $\theta_1 \approx \log(1001)$.

In case of misspecification, the implicit update will try to “overfit” on the current data point but does not diverge like explicit SGD.

BIAS

Let $\boldsymbol{J} = \psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)$. It holds:

$$||\mathrm{E}(\boldsymbol{\theta}_n^{\mathrm{sgd}}) - \boldsymbol{\theta}^*|| \propto \prod_i^n ||(\boldsymbol{I} - a_i \boldsymbol{J})||$$

$$||\mathrm{E}(\boldsymbol{\theta}_n^{\mathrm{im}}) - \boldsymbol{\theta}^*|| \propto \prod_i^n ||(\boldsymbol{I} + a_i \boldsymbol{J})^{-1}||$$

- For large enough n,

$$||(\boldsymbol{I} - a_n \boldsymbol{J})|| \leq ||(\boldsymbol{I} + a_n \boldsymbol{J})^{-1}||$$

  and so the explicit SGD is converging *faster*. However, the asymptotic rates are equal.
- The spectra of $(\boldsymbol{I} - \epsilon \boldsymbol{J})$ and $(\boldsymbol{I} + \epsilon \boldsymbol{J})^{-1}$ are crucial for their stability properties.

STABILITY

Recall that $\boldsymbol{J} = \psi \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) \geq 0$. For $\epsilon > 0$,

- The spectrum of $(\boldsymbol{I} - \epsilon \boldsymbol{J})$ is equal to $(1 - \epsilon \lambda_i(\boldsymbol{J}))$. For stability, we thus need $|1 - \epsilon \lambda_i(\boldsymbol{J})| < 1$. The explicit updates are *conditionally stable*.

- The spectrum of $(\boldsymbol{I} + \epsilon \boldsymbol{J})^{-1}$ is $(1 + \epsilon \lambda_i(\boldsymbol{J}))^{-1} < 1$ and thus, the implicit updates are *unconditionally stable*.

ASYMPTOTIC VARIANCE

Let $\mathrm{Var}\left(\boldsymbol{\theta}_n^{\mathrm{sgd}}\right) = \boldsymbol{\Sigma}_n^{\mathrm{sgd}}$ and $\mathrm{Var}\left(\boldsymbol{\theta}_n^{\mathrm{im}}\right) = \boldsymbol{\Sigma}_n^{\mathrm{im}}$.

Theorem 3. *The asympotic variance of the explicit SGD estimator is,*

$$n \cdot \boldsymbol{\Sigma}_n^{\mathrm{sgd}} \to \alpha^2\psi^2(2\alpha\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) - \boldsymbol{I})^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) \tag{8}$$

The asymptotic variance of the implicit SGD estimator is,

$$n \cdot \boldsymbol{\Sigma}_n^{\mathrm{im}} \to \alpha^2\psi^2(2\alpha\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) - \boldsymbol{I})^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) \tag{9}$$

Assuming convergence, both SGD methods have the same asymptotic efficiency.

ASYMPTOTIC VARIANCE (CONTINUED)

- It holds that

$$a^2\psi^2(2a\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) - \boldsymbol{I})^{-1}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) \geq \underbrace{\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)^{-1}}_{\text{MLE(theoretically optimal)}} \quad , \forall \alpha, \psi > 0$$

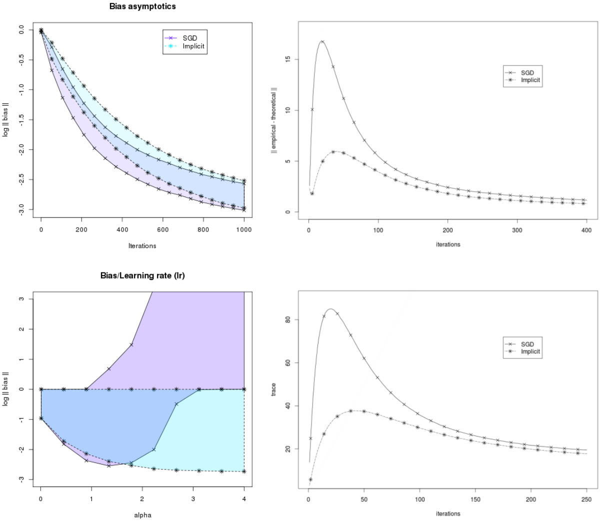
  Not surprisingly, both methods incur information loss.
- However, it is possible to **optimize** for the learning rate e.g., minimize the trace of the asymptotic variance:

$$\hat{\alpha} = \arg\min_a \sum_i \frac{a^2\lambda_i}{2a\lambda_i - 1} \tag{10}$$

  for $\lambda_i = \texttt{spectrum}(\psi\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*))$.

$$\boldsymbol{\theta}^* = \mathbf{1} \in \mathbb{R}^{20},\ \boldsymbol{x}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{V}),\ \lambda_i(\boldsymbol{V}) \in [0.2, 1],\ y_n \sim \mathcal{N}(\boldsymbol{x}_n^\intercal \boldsymbol{\theta}^*, 1).$$

(left) = log-bias (2.5-97.5% percentiles),
(right) trace of empirical covariance matrix over 2,000 samples



EXPERIMENTS ON CLASSIFICATION TASKS (SVM)

TESTING OUTSIDE GLM FAMILY

Table : Test errors of explicit and implicit SGD methods on the RCV1 dataset benchmark. Training times are roughly comparable. Best scores, for a particular loss and regularization, are bolded.

		REGULARIZATION (λ)		
LOSS		1E-5	1E-7	1E-12
HINGE	SGD	**4.65%**	**3.57%**	4.85%
	IMPLICIT	4.68%	3.6%	**3.46%**
LOG	SGD	5.23%	3.87%	5.42%
	IMPLICIT	**4.28%**	**3.69%**	**4.01%**

SUMMARY

- Exact statistical analysis of SGD is possible in GLMs, both for explicit and implicit updates.
- Helps in optimizing for the learning rate.
- Implicit updates compare favorably to explicit ones:
  - They are easy to implement (Theorem 1).
  - They have the same asymptotic performance (bias and variance, Theorems 2,3).
  - They are unconditionally stable, and thus more robust to misspecification.
  - Vanilla implementation performs on par with standard SGD on large-scale optimization tasks.

Thank you!

(poster #T-76).