Rapid Stochastic Gradient Descent for Atomic Learning

Nicol N. Schraudolph
Why Stochastic Gradient?
The flood of information caused by
- plentiful, affordable sensors (such as webcams)
- ever-increasing networking of these sensors

overwhelms our processing ability in, e.g.,
- **science** - pulsar survey at Arecibo: 1 TB/day
- **business** - Dell website: over 100 page requests/sec
- **security** - London: over 500,000 security cameras

We need intelligent, adaptive filters to cope!
The imagination driving Australia’s ICT future.

The Challenge for ML

Coping with the info glut requires algorithms for

- large adaptive models
  - millions of degrees of freedom
- large volumes of low-quality data
  - noisy, correlated, non-stationary, outliers
- efficient real-time, online adaptation
  - no fixed training set, life-long learning

Current optimization techniques are inadequate.
Online Learning Paradigm

classical optimization:

iterative optimizer

objective fn.

training data set

nested loops!

online learning:

online optimizer

training data stream

(aka adaptive filtering, stochastic approximation, ...)

Statistical Machine Learning Program
Stochastic Approximation

Classical formulation of optimization problem:

\[ \theta^* = \arg\min_{\theta} : \quad E_x [J(\theta, x)] \approx \frac{1}{|X|} \sum_{x_i \in X} J(\theta, x_i) \]

- inefficient for large data sets \( X \)
- inappropriate for never-ending, potentially non-stationary data streams

\[ \Rightarrow \text{must resort to stochastic approximation:} \]

\[ \theta_{t+1} \approx \arg\min_{\theta} J(\theta, x_t) \quad (t = 0, 1, 2, \ldots) \]
Stochastic Objective
The Key Problem

Online, scalable optimization algorithms:
- evolutionary algorithms
- gradient descent
- conjugate gradient
- quasi-Newton
- Kalman Filter
- Levenberg Marquardt

Convergence speed:
- $O(1)$
- $O(n)$
- $O(n^2)$
- $O(n^3)$

Cost per iteration:
- online, scalable

Statistical Machine Learning Program
www.nicta.com.au
The Key Problem

Stochastic approximation breaks many optimizers:

- conjugate directions break down due to noise
- line minimizations (CG, quasi-Newton) inaccurate
- Newton, Levenberg-Marquardt, Kalman filter - too expensive for large-scale problems

This only leaves

- evolutionary alg.s - very inefficient (don’t use gradient)
- simple gradient descent - can be slow to converge
Stochastic Meta-Descent (SMD)
Gain Vector Adaptation

Given stochastic gradient \( g_t := \partial_{\theta} J(\theta_t, x_t) \), adapt \( \theta \) by gradient descent with gain vector \( \eta \):

\[
\theta_{t+1} = \theta_t - \eta_t \cdot g_t
\]

Key idea:

simultaneously adapt \( \eta \) by exponentiated gradient:

\[
\ln \eta_t = \ln \eta_{t-1} - \mu \partial_{\ln \eta} J(\theta_t, x_t)
\]

\[
\eta_t = \eta_{t-1} \cdot \exp(-\mu \partial_{\theta} J(\theta_t, x_t) \cdot \partial_{\ln \eta} \theta_t)
\]

\[
\approx \eta_{t-1} \cdot \max\left(\frac{1}{2}, 1 - \mu g_t \cdot v_t\right)
\]
Conventionally, \( v_{t+1} := \partial_{\ln \eta_t} \theta_{t+1} = -\eta_t \cdot g_t \)

(recall that \( \theta_{t+1} = \theta_t - \eta_t \cdot g_t \) )

giving \( \eta_t = \eta_{t-1} \cdot \max(\frac{1}{2}, 1 + \mu \eta_{t-1} \cdot g_{t-1} \cdot g_t) \)

⇒ adaptation of \( \eta \) driven by autocorrelation of \( g \)
To capture long-term dependence of $\theta$ on $\eta$:

$$
\nu_{t+1} := \sum_{i=0}^{t} \lambda^i \frac{\partial \theta_{t+1}}{\partial \ln \eta_{t-i}}
$$

define $\nu_{t+1}$ with decay $0 \leq \lambda \leq 1$ (free parameter)
The imagination driving Australia’s ICT future.

SMD’s $\mathbf{v}$-update

\[ \mathbf{v}_{t+1} = \lambda \mathbf{v}_t - \mathbf{\eta}_t \cdot (\mathbf{g}_t + \lambda \mathbf{H}_t \mathbf{v}_t) \]

- we obtain a simple iterative update for $\mathbf{v}$
- $\lambda$ can smoothe over correlated input signals
- iteration similar to TD($\lambda$) RL method (Sutton)
- generalizes Sutton’s (1992) KL algorithm
  linear to non-linear system, diagonal to full Hessian
Fixpoint of $\mathbf{v}$

Fixpoint of

$$v_{t+1} = \lambda v_t - \eta_t \cdot (g_t + \lambda H_t v_t)$$

is a Levenberg-Marquardt style gradient step:

$$\mathbf{v} \rightarrow -[\lambda H + (1 - \lambda)\text{diag}(\eta)^{-1}]^{-1}g$$

- $\mathbf{v}$ is too noisy to use directly; SMD achieves stability by means of the double integration $\mathbf{v} \rightarrow \eta \rightarrow \theta$
- $\mathbf{v} \cdot g$ is well-behaved (self-normalizing property)
- non-convex fn.s: use Gauss-Newton approximation
Fast $HV$ Product

Explicit computation of $HV$ product would be $O(n^2)$. 

**But:** consider differential

$$dg(\theta) = H(\theta) \, d\theta$$

- can set $d\theta := v$, forward-propagate through $g()$
- as **efficient** as 2-3 gradient evaluations (typ. $O(n)$)
- matched approximations of $g$ and $HV$ $\Rightarrow$ robust
- can even co-opt **complex** arithmetic:

$$g(\theta + i\epsilon d\theta) = g(\theta) + O(\epsilon^2) + i\epsilon dg(\theta) \quad (\epsilon \approx 10^{-150})$$
SMD Benchmarks
Four Regions Benchmark

Compare simple stochastic gradient (**SGD**), conventional gain vector adaptation (**ALAP**), stochastic meta-descent (**SMD**), and a global extended Kalman filter (**GEKF**).
Benchmark: Convergence

The imagination driving Australia’s ICT future.
### Computational Cost

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>storage</th>
<th>weight</th>
<th>flops</th>
<th>update</th>
<th>CPU ms</th>
<th>pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>SMD</td>
<td>3</td>
<td>18</td>
<td></td>
<td></td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>ALAP</td>
<td>4</td>
<td>18</td>
<td></td>
<td></td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>GEKF</td>
<td>&gt;90</td>
<td>&gt;1500</td>
<td></td>
<td></td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>
Benchmark: CPU Usage
Autocorrelated Data

- i.i.d. uniform
- Sobol
- Brownian

Patterns

- vario-eta
- ALAP
- ELK1
- SMD

Mom.

\[ E \]

Comparison to CG

Conjugate Gradient
- deterministic
- (1000 pts)
- overfits

Stochastic Model Drift (SMD)
- stochastic
- (1000 pts/iteration)
- diverges

- stochastic
- (5 pts/iteration)
- converges
SMD Applications
Application: Turbulent Flow

(PhD thesis of M. Milano, Inst. of Computational Science, ETH Zürich)

original flow
(75’000 d.o.f.)

linear PCA
(160 p.c.)

neural network
(160 nonlinear p.c.)
Hand Tracking with SMD

(PhD thesis of M. Bray, L. van Gool’s Computer Vision Lab, ETH Zürich)

Annealed Particle Filter
(114 sec/frame)

Stochastic Meta-Descent
(3 sec/frame)
Hand Tracking @ NICTA

with Desmond Chik (PhD) and Jochen Trumpf (SEACS)

- detailed hand model (26 dof, ~10k vertices, skin blending, ...)
- randomly sample few points on model surface, project to image
- compare camera image(s) there, use resulting stochastic gradient to adjust model
- radical: completely marker- and feature-free tracking
SMD & Policy Gradient RL

with Jin Yu (PhD) and Doug Aberdeen (SML)

- SMD accelerates PG-RL
- complex interaction with temporal task structure
- had a poster at NIPS’05
SMD for Online SVM

Online SVM aka NORMA (Kivinen, Smola, Williamson 2004):
- online kernel method
- stochastic gradient in expansion coefficients
- employs scalar gain $\eta$

Applied SMD (with Vishy):
- $\mathbf{v}$ is function in RKHS
- $\langle \mathbf{g}, \mathbf{v} \rangle$ can be maintained incrementally in $O(n)$
- presented at NIPS’05 workshop, submitted to JMLR
Training CRFs with SMD

- 1-D CRF chain for ConNLL-2000 Base NP chunking
- (predictable!) huge speed-up for online learning
Training CRFs with SMD

- BioNLP/NLPBA-2004 named-entity recognition task
- 2-D CRF lattices for vision (M. Schmidt & K. Murphy, UBC) work well with loopy BP; more robust to overfitting
SMD on 2-D CRF Lattices

original    log. regr.    BFGS    SGD    SMD    SMD (PL)
ground truth logistic regression SMD (loopy BP)
Summary:
- data-rich ML problems need stochastic approximation
- classical gradient methods are not up to the task
- SMD provides gain adaptation for stochastic gradient
  \((Hv \text{ product gives cheap second-order information})\)

Wish List:
- convergence & stability analysis for SMD (volunteers?)
- gain matrix version of SMD (rotation invariance)
- online LBFGS; proof that online CG can’t work