Bridging the Gap between Stochastic Gradient MCMC and Stochastic Optimization

Changyou Chen[1], David Carlson[2], Zhe Gan[1], Chunyuan Li[1] and Lawrence Carin[1]

[1] Duke University, Durham, NC [2] Columbia University, New York, NY

INTRODUCTION

Problem of interest: How to better learn a complex and high-dimensional model in a big-data setting

Motivation:

- Combine approaches and advantages from stochastic gradient MCMC (SG-MCMC) and stochastic optimization
- Stochastic optimization:
  - computationally efficient iterations, fast convergence to a local optima
- Stochastic gradient MCMC:
  - computationally efficient iterations, slower convergence, able to explore the parameter space

Main idea: Begin with a preconditioned SG-MCMC algorithm, and gradually anneal the system temperature to zero such that it becomes a preconditioned stochastic optimization algorithm

Advantages: Our algorithm (Santa) has both an adaptive preconditioner and adaptive momentum, not available in existing algorithms.

Software: Code available at https://github.com/cchangyou/Santa.

STOCHASTIC OPTIMIZATION vs. SG-MCMC

Stochastic optimization

- Stochastic gradient descent (SGD)
  - basic stochastic optimization algorithm, without momentum and preconditioning
- SGD with momentum (SGD-M)
  - extending SGD with momentum
- RMSProp, ADAprop, Adadelta, ...
  - extends SGD with a preconditioner
- Adam
  - extending SGD with both momentum and preconditioning

Stochastic gradient MCMC

- Stochastic gradient Langevin dynamics (SGLD)
  - Sampling analog of SGD, without momentum and preconditioning
- Stochastic gradient Hamiltonian Monte Carlo (SGHMC)
  - Sampling analog of SGD-M, with momentum
- Preconditioned stochastic gradient Langevin dynamics (pSGLD)
  - Sampling analog of RMSProp/Adagrad, with a preconditioner
- Multivariate stochastic gradient thermostats (mSGNHT)
  - Sampling with element-wise adaptive momentum, no obvious stochastic optimization analog

Table: SG-MCMC algorithms and stochastic optimization analags

Algorithms	SG-MCMC		Optimization
Basic	SGLD	$\Longleftrightarrow$	SGD
Precondition	pSGLD	$\Longleftrightarrow$	RMSprop
Momentum	SGHMC	$\Longleftrightarrow$	SGD-M
Thermostat	mSGNHT	$\Longleftrightarrow$	Santa

THE SANTA ALGORITHM

Input: η_t (learning rate), σ, λ, $burnin$, $\beta = \{\beta_1, \beta_2, \cdots\} \to \infty$, $\{\boldsymbol{\zeta}_t \in \mathbb{R}^p\} \sim N(\mathbf{0}, \mathbf{I}_p)$.
Initialize $\boldsymbol{\theta}_0$, $\boldsymbol{u}_0 = \sqrt{\eta} \times N(0, I)$, $\boldsymbol{\alpha}_0 = \sqrt{\eta}C$, $\boldsymbol{v}_0 = 0$.
for $t = 1, 2, \ldots$ **do**
  Evaluate $\tilde{\boldsymbol{f}}_t \triangleq \nabla_{\boldsymbol{\theta}} \tilde{U}(\boldsymbol{\theta}_{t-1})$ on the t^{th} mini-batch.
  $\boldsymbol{v}_t = \sigma \boldsymbol{v}_{t-1} + \frac{1-\sigma}{N^2} \tilde{\boldsymbol{f}}_t \odot \tilde{\boldsymbol{f}}_t$.
  $\boldsymbol{g}_t = 1 \oslash \sqrt{\lambda + \sqrt{\boldsymbol{v}_t}}$.
  if $t < burnin$ **then**
    $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + (\boldsymbol{u}_{t-1} \odot \boldsymbol{u}_{t-1} - \eta/\beta_t)$. # exploration
    $\boldsymbol{u}_t = \frac{\eta}{\beta_t}(1 - \boldsymbol{g}_{t-1} \oslash \boldsymbol{g}_t) \oslash \boldsymbol{u}_{t-1} + \sqrt{\frac{2\eta}{\beta_t}} \boldsymbol{g}_{t-1} \odot \boldsymbol{\zeta}_t$.
  else
    $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}$. # refinement
    $\boldsymbol{u}_t = \mathbf{0}$.
  end if
  $\boldsymbol{u}_t = \boldsymbol{u}_t + (1 - \boldsymbol{\alpha}_t) \odot \boldsymbol{u}_{t-1} - \eta \boldsymbol{g}_t \odot \tilde{\boldsymbol{f}}_t$.
  $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{g}_t \odot \boldsymbol{u}_t$.
end for

THEORY

- The Santa algorithm is based on the following stochastic differential equations, whose marginal distribution corresponds to the true posterior distribution of interest at temperature $\frac{1}{\beta}$

$$\begin{cases} \mathrm{d}\boldsymbol{\theta} = G_1(\boldsymbol{\theta})\boldsymbol{p}\mathrm{d}t \\ \mathrm{d}\boldsymbol{p} = \big(-G_1(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) - \Xi\boldsymbol{p} + \frac{1}{\beta}\nabla_{\boldsymbol{\theta}} G_1(\boldsymbol{\theta}) \\ \quad\quad + G_1(\boldsymbol{\theta})(\Xi - G_2(\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}} G_2(\boldsymbol{\theta})\big)\,\mathrm{d}t + (\frac{2}{\beta}G_2(\boldsymbol{\theta}))^{\frac{1}{2}}\mathrm{d}w \\ \mathrm{d}\Xi = \big(\boldsymbol{Q} - \frac{1}{\beta}I\big)\,\mathrm{d}t\ , \end{cases} \quad (1)$$

where $\boldsymbol{Q} = \mathrm{diag}(\boldsymbol{p} \odot \boldsymbol{p})$, w is standard Brownian motion, $\boldsymbol{G}_1(\boldsymbol{\theta})$ and $\boldsymbol{G}_2(\boldsymbol{\theta})$ are preconditioners.

- Santa algorithm is derived by solving (1) numerically with an increasing sequence of β

Theorem (Convergence)

With certain assumptions, the Santa algorithm converges in expectation to a global optima of a smooth function.

EXPERIMENTS

Illustration: Double-well potential example.

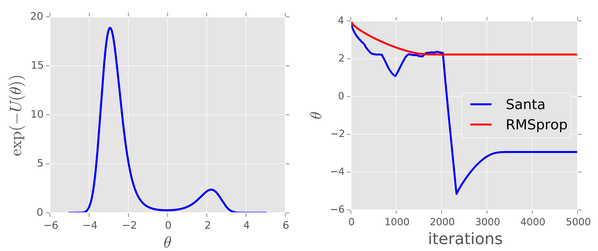

Figure: (Left) Double-well potential. (Right) The evolution of θ using Santa and RMSprop algorithms.

Feedforward neural networks and convolutional neural networks

- Santa outperforms other algorithms in most cases.

Table: Test error on MNIST classification using FNN and CNN.

Algorithms	FNN-400	FNN-800	CNN
Santa	**1.21%**	**1.16%**	**0.47%**
Adam	1.53%	1.47%	0.59%
RMSprop	1.59%	1.43%	0.64%
SGD-M	1.66%	1.72%	0.77%
SGD	1.72%	1.47%	0.81%
SGLD	1.64%	1.41%	0.71%
BPB$^\diamond$	1.32%	1.34%	–
SGD, Dropout$^\diamond$	1.51%	1.33%	–
Stoc. Pooling$^\triangleright$	–	–	0.47%
NIN, Dropout$^\circ$	–	–	0.47%
Maxout, Dropout*	–	–	0.45%

Recurrent neural networks

- Language modeling with an RNN; Tests on four publicly available datasets.

Table: Test negative log-likelihood on 4 datasets.

Algorithms	Piano.	Nott.	Muse.	JSB.
Santa	**7.60**	**3.39**	**7.20**	**8.46**
Adam	8.00	3.70	7.56	8.51
RMSprop	7.70	3.48	7.22	8.52
SGD-M	8.32	3.60	7.69	8.59
SGD	11.13	5.26	10.08	10.81
HF$^\diamond$	7.66	3.89	**7.19**	8.58
SGD-M$^\diamond$	8.37	4.46	8.13	8.71

GoogleNet for ImageNet classification

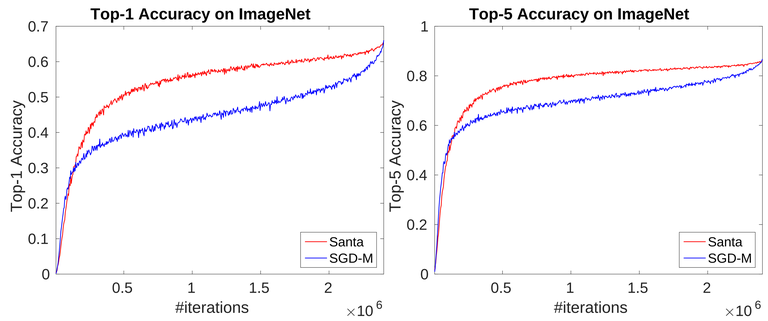

Figure: Santa vs. SGD with momentum on ImageNet.

ACKNOWLEDGEMENTS

This research was supported by ARO, DARPA, DOE, NGA, ONR and NSF.