STOCHASTIC OPTIMIZATION IN MACHINE LEARNING

Case Studies in Nonlinear Optimization

F. Bauer S. Chambon R. Halbig S. Heidekrüger J. Heuke

July 6, 2015

Technische Universität München

We're not running out of data anytime soon. It's maybe the only resource that grows exponentially.

Andreas Weigend

OUTLINE

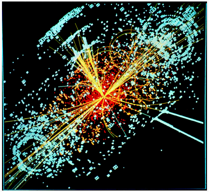
INTRODUCTION

Implementation of autonomously learning software for:

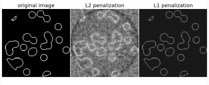
- Discovery of patterns and relationships in data
- Prediction of future events

Examples:

Higgs-Boson



Compressed Sensing



EEG

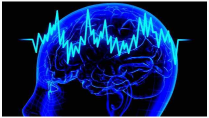


Image Reconstruction

CHALLENGES IN MACHINE LEARNING

- Massive amounts of training data
- Construction of very large models
- Handling high memory/computational demands

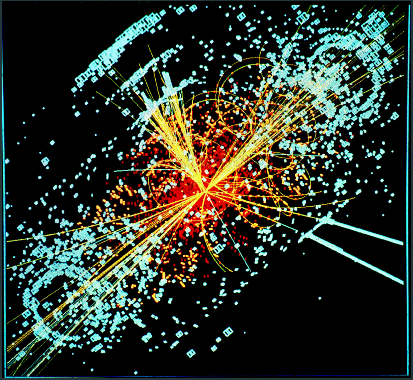
Stochastic Methods

STOCHASTIC FRAMEWORK

A STOCHASTIC QUASI-NEWTON METHOD

Classification

Did we just detect a Higgs-Boson?



What is it? Why? Main ideas, high-level pseudo code overview? short bfgs repitition? Extreme Cases (L-BFGS, SGD)

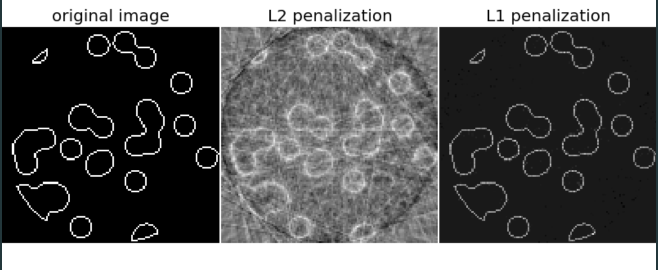
Explain the Dataset quickly. Why is this good for SQN testing? Why is it challenging? (file size etc)

Pretty picures about the behaviour of SQN on HIGGS and comparison with traditional SGD

PROXIMAL METHOD

Image Reconstruction

What did the original image look like?



Problem

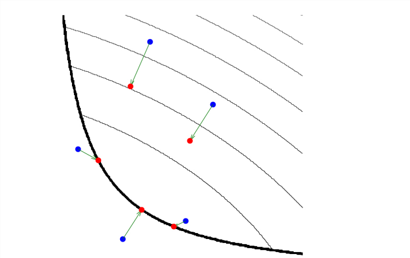
$$\min_{x} F(x) := \underbrace{f(x)}_{smooth} + \underbrace{h(x)}_{non-smooth}$$

Problem

$$\min_x F(x) := \underbrace{f(x)}_{smooth} + \underbrace{h(x)}_{non-smooth}$$

Proximity Operator

$$\mathrm{prox}_f(v) = \operatorname*{argmin}_x \left(f(x) + \frac{1}{2}\|x - v\|_2^2\right)$$



Traditional Proximal Gradient Step:

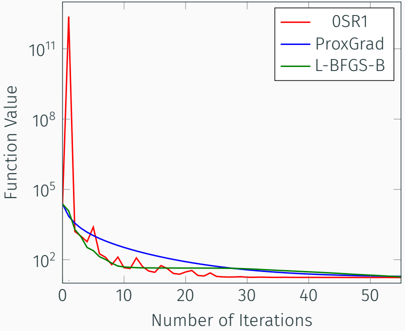
$$x_{k+1} = \mathrm{prox}_{\lambda_k h}(x_k - \lambda_k \nabla f(x_k))$$

Quasi-Newton Proximal Step:

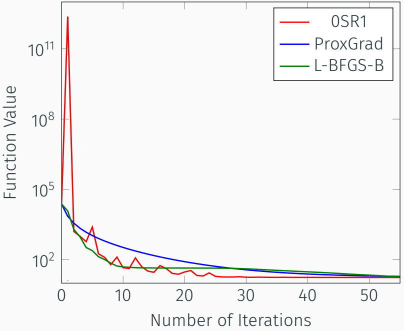
$$x_{k+1} = \mathrm{prox}_h^{B_k}(x_k - B_k^{-1} \nabla f(x_k)),$$

with $B_k = \underbrace{D_k}_{diag} + \underbrace{u_k}_{\in \mathbb{R}^n} u_k^T$.

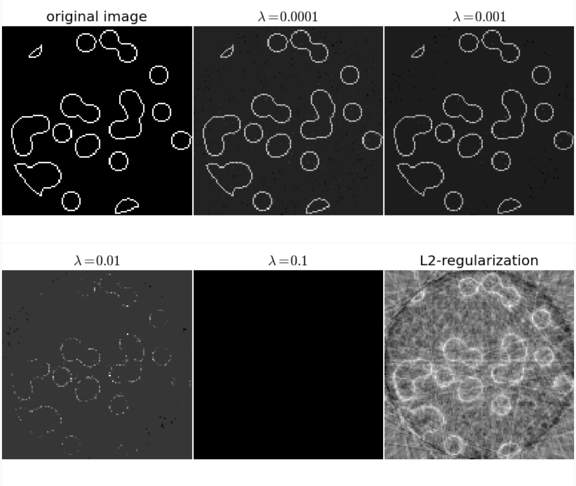
$$F(x) = \|Ax - b\| + \lambda\|x\|_1$$
$$A \in \mathbb{R}^{1500\times 3000},\ b \in \mathbb{R}^{1500}$$
$$A_{ij},\ b_i \sim \mathcal{N}(0,1),\ \lambda = 0.1$$



$$F(x) = \|Ax - b\| + \lambda\|x\|_1$$
$$A \in \mathbb{R}^{1500\times 3000},\ b \in \mathbb{R}^{1500}$$
$$A_{ij},\ b_i \sim \mathcal{N}(0,1),\ \lambda = 0.1$$



Effect of regularization parameter λ on solution:



LOGISTIC REGRESSION: AN EXAMPLE

Explain what we want to do, and explain the dataset, and why using both SQN and Prox makes sense

Nice table with SQN, SGD (no reg, L2), (Lasso,) Prox (L1) showing Obj. value in found optimum, CPU time, Iterations, F1 score of prediction model

Use different reg. parameters?? Stop after fixed time? after fixed iters? after insign. improvements

CONCLUSION

SUMMARY

QUESTIONS?

S. Becker and J. Fadili.
A quasi-newton proximal splitting method.
In *Advances in Neural Information Processing Systems*, pages 2618–2626, 2012.