The imagination driving Australia's ICT future.

NATIONAL ICT AUSTRALIA

Nicol N. Schraudolph

Rapid Stochastic Gradient Descent for Atomic Learning

PASCAL
Pattern Analysis, Statistical Modelling and Computational Learning

Australian Government
Department of Communications, Information Technology and the Arts
Australian Research Council

NICTA Members

ANU THE AUSTRALIAN NATIONAL UNIVERSITY · UNSW · NSW First for Business · Department of State and Regional Development · BusinessACT ACT GOVERNMENT

Victoria The Place To Be · THE UNIVERSITY OF MELBOURNE · The University of Sydney · Queensland Government · Griffith UNIVERSITY · QUT Queensland University of Technology · THE UNIVERSITY OF QUEENSLAND AUSTRALIA

NICTA Partners

Why Stochastic Gradient?

The Information Glut

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 Challenge for ML

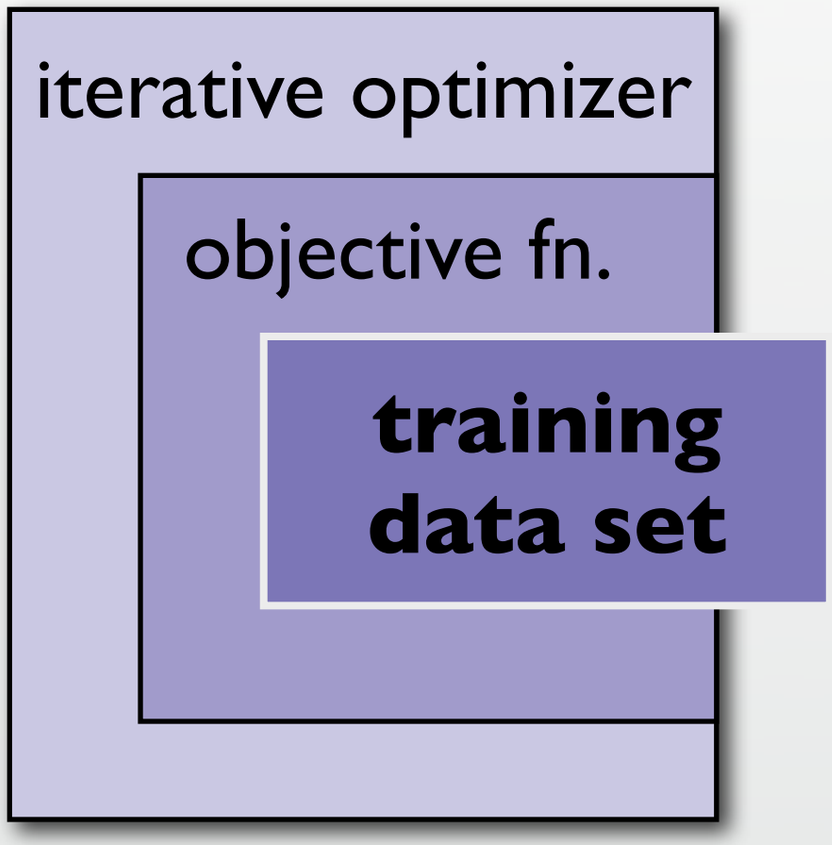
Coping with the info glut requires algorithms for

- 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.

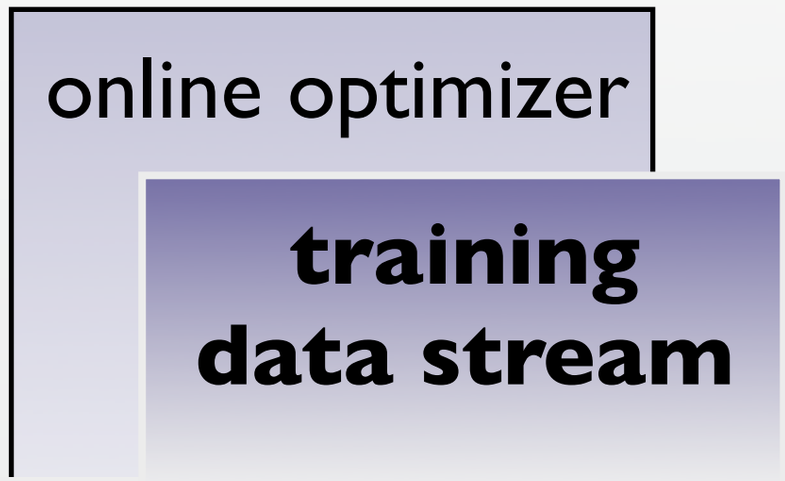
Online Learning Paradigm

classical optimization:



nested loops!

online learning:



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

Stochastic Approximation

Classical formulation of optimization problem:

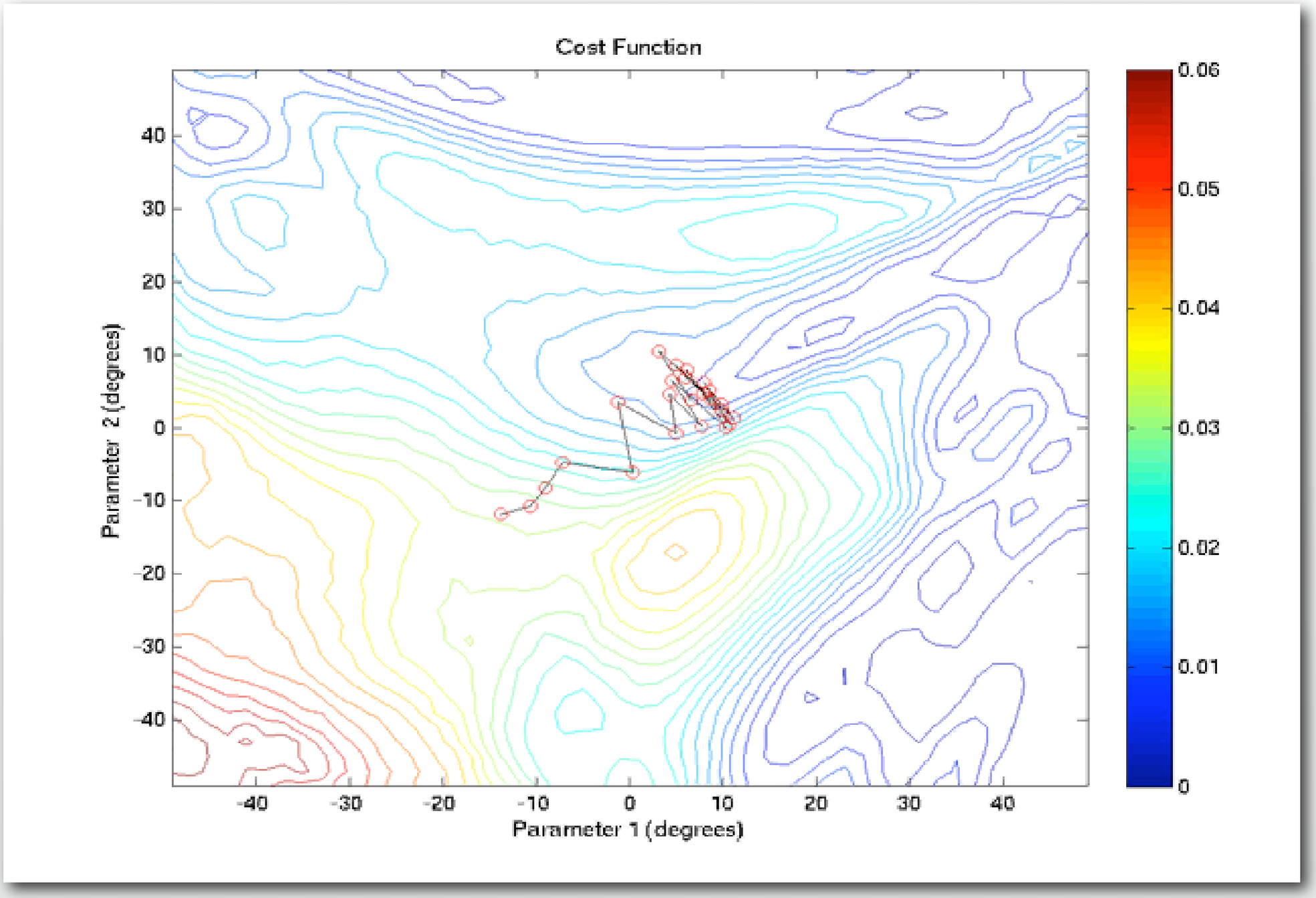
$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} : \quad E_{\boldsymbol{x}}[J(\boldsymbol{\theta}, \boldsymbol{x})] \approx \frac{1}{|X|} \sum_{\boldsymbol{x}_i \in X} J(\boldsymbol{\theta}, \boldsymbol{x}_i)$$

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

$\Rightarrow$ must resort to stochastic approximation:

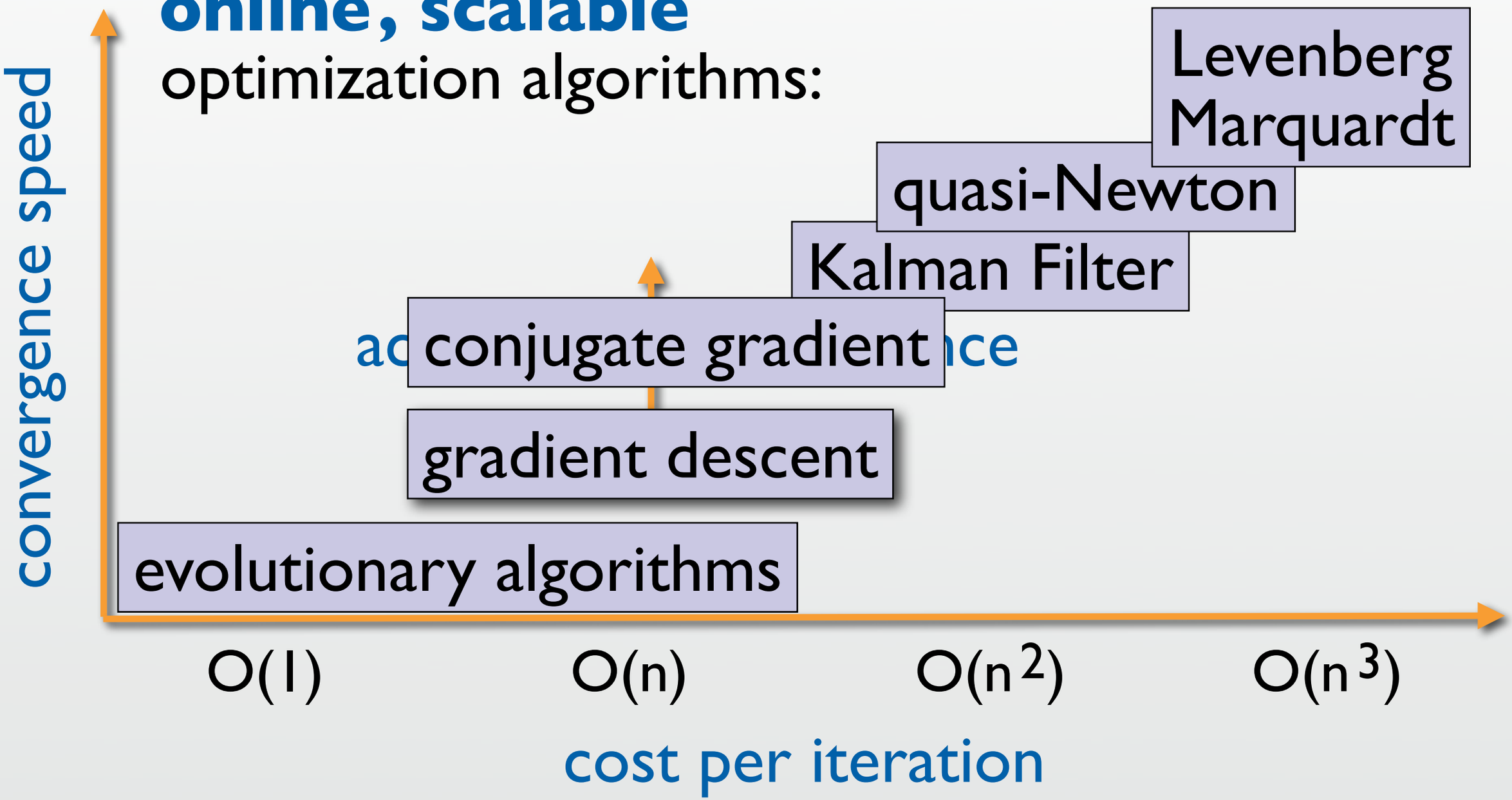
$$\boldsymbol{\theta}_{t+1} \approx \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t, \boldsymbol{x}_t) \qquad (t = 0, 1, 2, \ldots)$$

Stochastic Objective



The Key Problem

online, scalable optimization algorithms:



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)

NATIONAL ICT AUSTRALIA

Gain Vector Adaptation

Given stochastic gradient $g_t := \partial_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t, \boldsymbol{x}_t)$,
adapt $\boldsymbol{\theta}$ by gradient descent with gain vector $\boldsymbol{\eta}$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \boldsymbol{\eta}_t \cdot \boldsymbol{g}_t$$

Key idea:
simultaneously adapt $\boldsymbol{\eta}$ by exponentiated gradient:

$$\ln \boldsymbol{\eta}_t = \ln \boldsymbol{\eta}_{t-1} - \mu \, \partial_{\ln \boldsymbol{\eta}} J(\boldsymbol{\theta}_t, \boldsymbol{x}_t)$$

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_{t-1} \cdot \exp(-\mu \, \partial_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t, \boldsymbol{x}_t) \cdot \partial_{\ln \boldsymbol{\eta}} \boldsymbol{\theta}_t)$$

$$\approx \boldsymbol{\eta}_{t-1} \cdot \max(\tfrac{1}{2}, 1 - \mu \, \boldsymbol{g}_t \cdot \boldsymbol{v}_t)$$

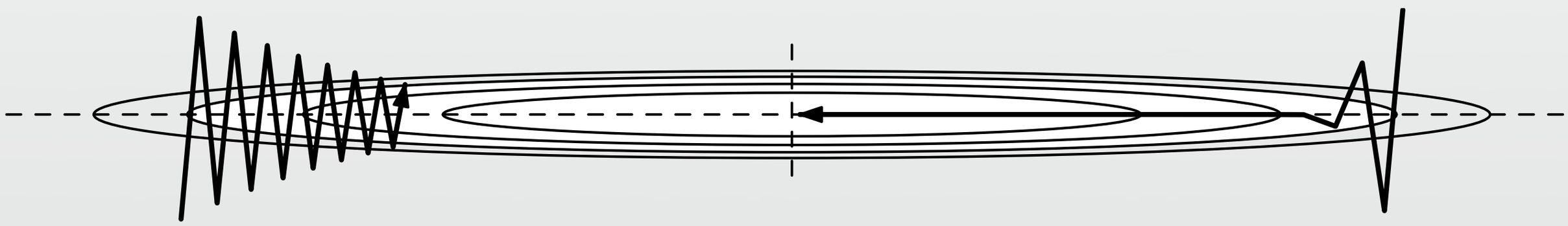
Single-Step Model

Conventionally, $\boldsymbol{v}_{t+1} := \partial_{\ln \boldsymbol{\eta}_t} \boldsymbol{\theta}_{t+1} = -\boldsymbol{\eta}_t \cdot \boldsymbol{g}_t$

(recall that $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \boldsymbol{\eta}_t \cdot \boldsymbol{g}_t$)

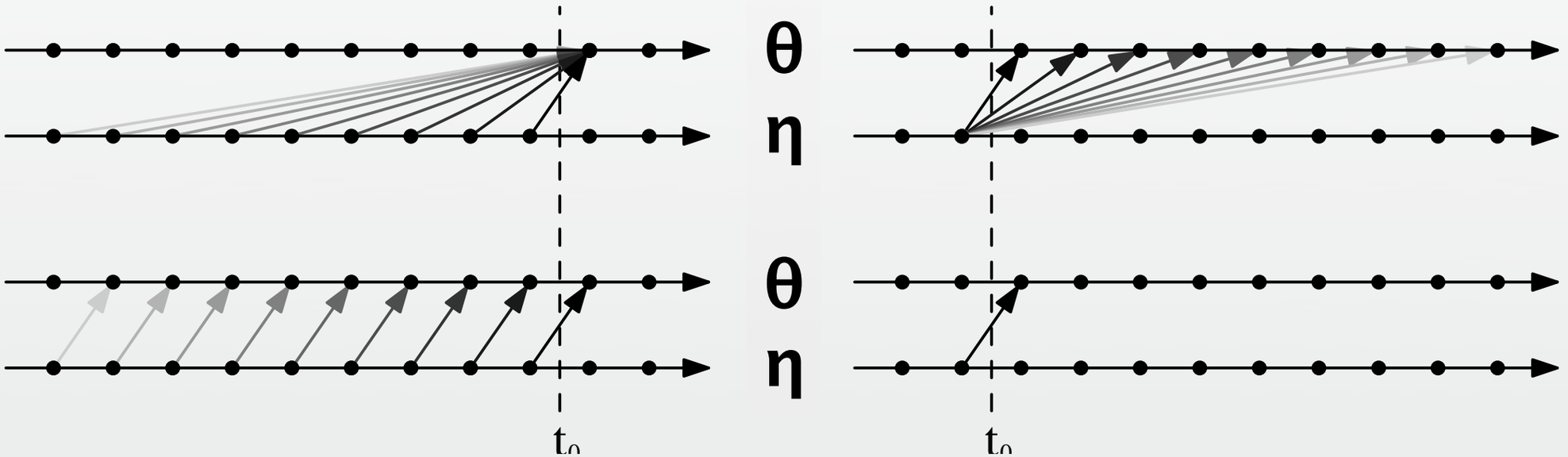
giving $\boldsymbol{\eta}_t = \boldsymbol{\eta}_{t-1} \cdot \max(\frac{1}{2}, 1 + \mu\, \boldsymbol{\eta}_{t-1} \cdot \boldsymbol{g}_{t-1} \cdot \boldsymbol{g}_t)$

⇒ adaptation of **η** driven by autocorrelation of **g**:



SMD's Multi-Step Model

To capture long-term dependence of θ on η:



define $\boldsymbol{v}_{t+1} := \sum_{i=0}^{t} \lambda^i \frac{\partial \boldsymbol{\theta}_{t+1}}{\partial \ln \boldsymbol{\eta}_{t-i}}$

decay $0 \leq \lambda \leq 1$
(free parameter)

SMD's **v**-update

$$\boldsymbol{v}_{t+1} = \lambda \boldsymbol{v}_t - \boldsymbol{\eta}_t \cdot (\boldsymbol{g}_t + \lambda \boldsymbol{H}_t \boldsymbol{v}_t)$$

- we obtain a simple iterative update for **v**
- λ can smoothe over correlated input signals
- iteration similar to TD(λ) RL method (Sutton)
- generalizes Sutton's (1992) K1 algorithm
  linear to non-linear system, diagonal to full Hessian

Fixpoint of $\mathbf{v}$

Fixpoint of $\boldsymbol{v}_{t+1} = \lambda \boldsymbol{v}_t - \boldsymbol{\eta}_t \cdot (\boldsymbol{g}_t + \lambda \boldsymbol{H}_t \boldsymbol{v}_t)$

is a Levenberg-Marquardt style gradient step:

$$\boldsymbol{v} \;\rightarrow\; -[\lambda \boldsymbol{H} + (1-\lambda)\mathrm{diag}(\boldsymbol{\eta})^{-1}]^{-1}\boldsymbol{g}$$

- $\mathbf{v}$ is too noisy to use directly; SMD achieves stability by means of the double integration $\mathbf{v} \rightarrow \boldsymbol{\eta} \rightarrow \boldsymbol{\theta}$
- $\mathbf{v} \cdot \mathbf{g}$ 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(n²).
But: consider differential

$$\mathrm{d}\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

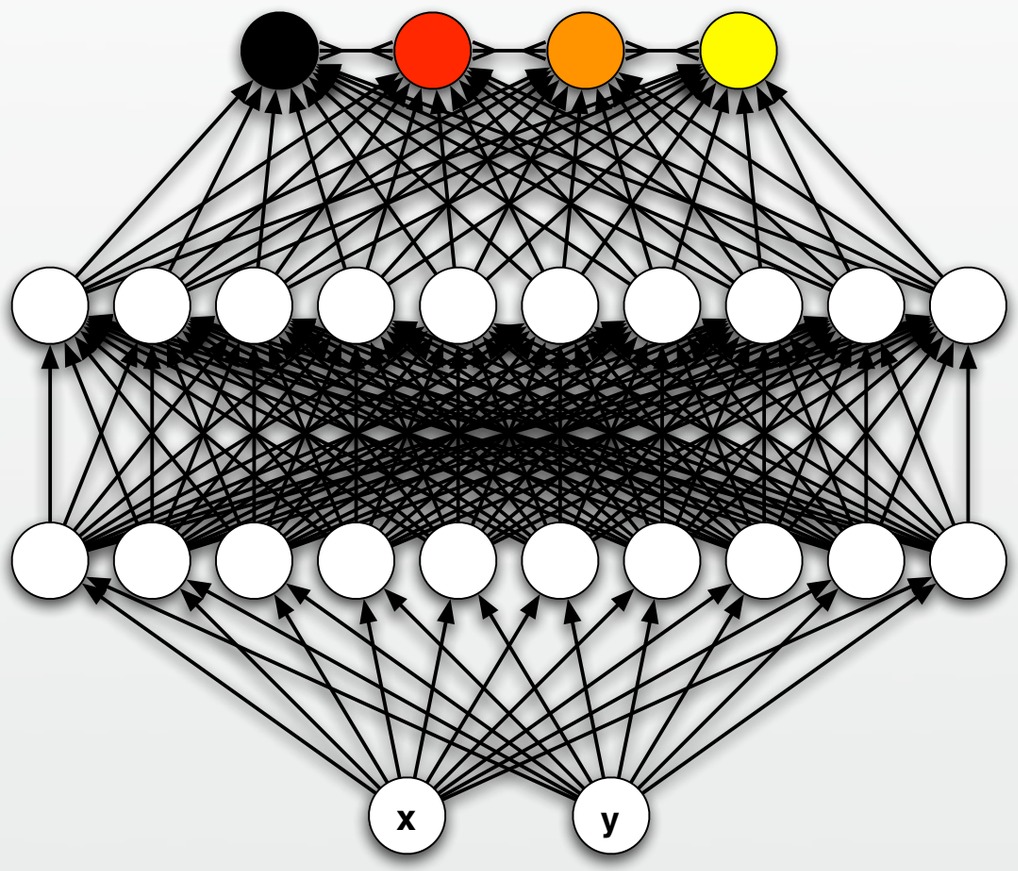
- can set d**θ** := **v**, forward-propagate through **g**()
- as efficient as 2-3 gradient evaluations (typ. O(n))
- matched approximations of **g** and **Hv** ⇒ robust
- can even co-opt complex arithmetic:

$$\boldsymbol{g}(\boldsymbol{\theta} + i\epsilon\mathrm{d}\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{\theta}) + O(\epsilon^2) + i\epsilon\mathrm{d}\boldsymbol{g}(\boldsymbol{\theta}) \qquad (\epsilon \approx 10^{-150})$$

SMD

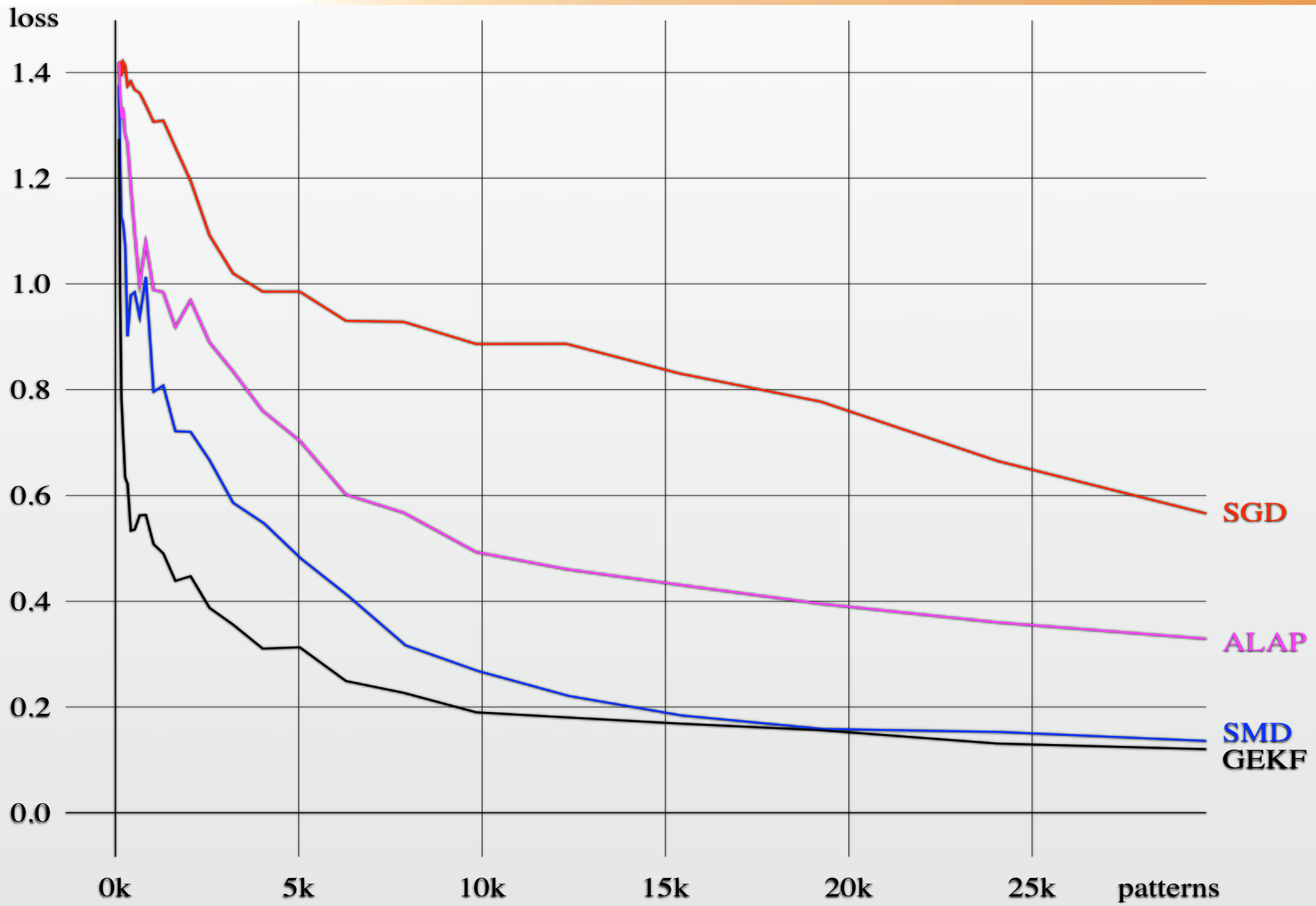
Benchmarks

Four Regions Benchmark



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

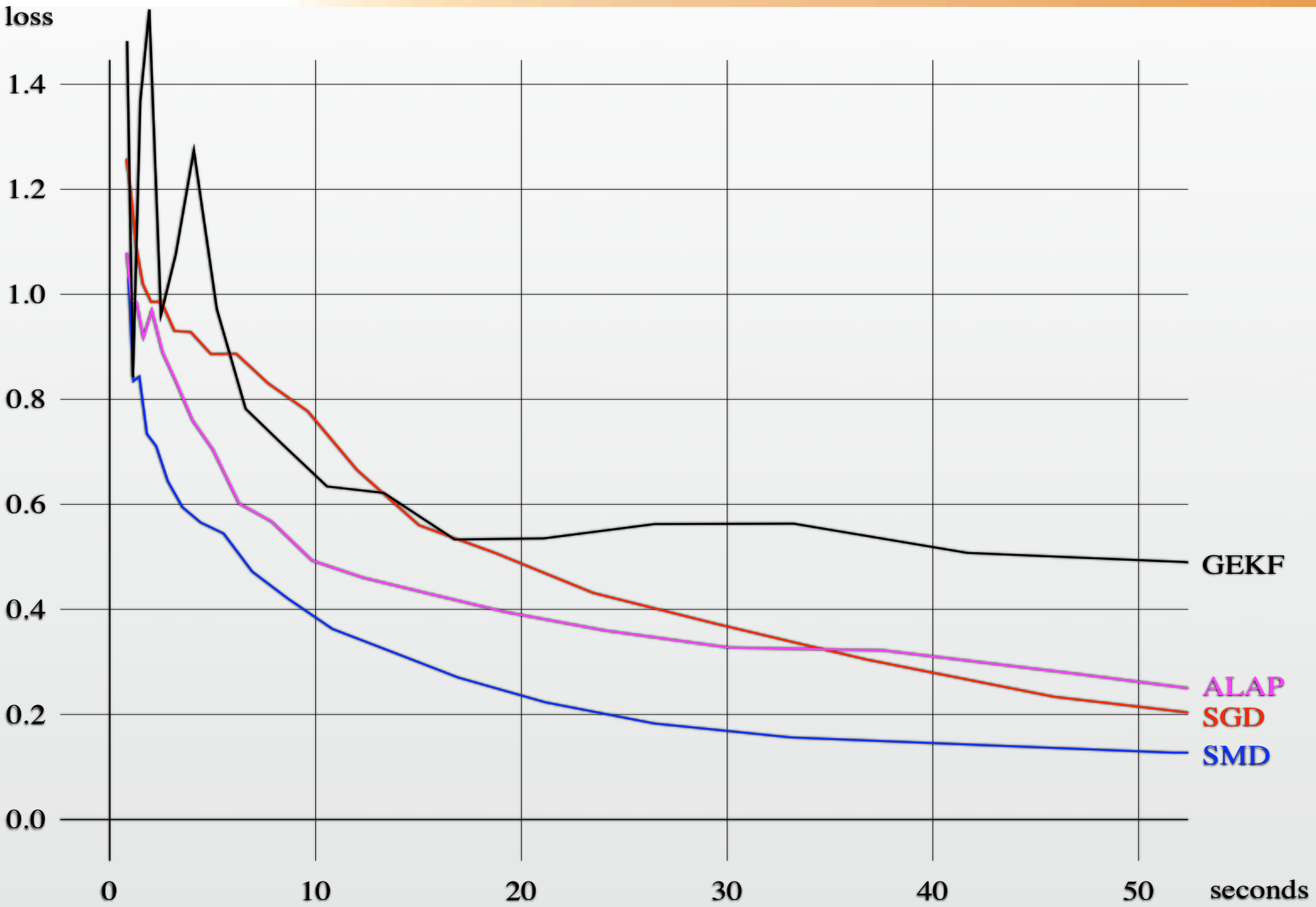
Benchmark: Convergence



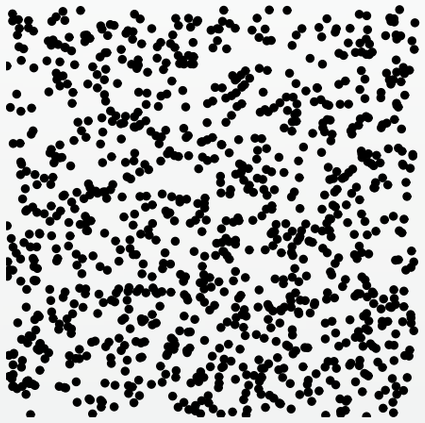
Computational Cost

Algorithm	storage weight	flops update	CPU ms pattern
SGD	1	6	0.5
SMD	3	18	1.0
ALAP	4	18	1.0
GEKF	>90	>1500	40

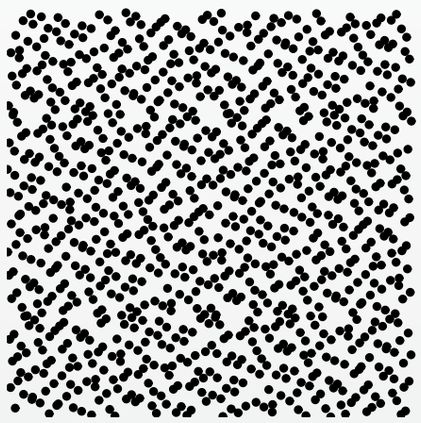
Benchmark: CPU Usage



Autocorrelated Data

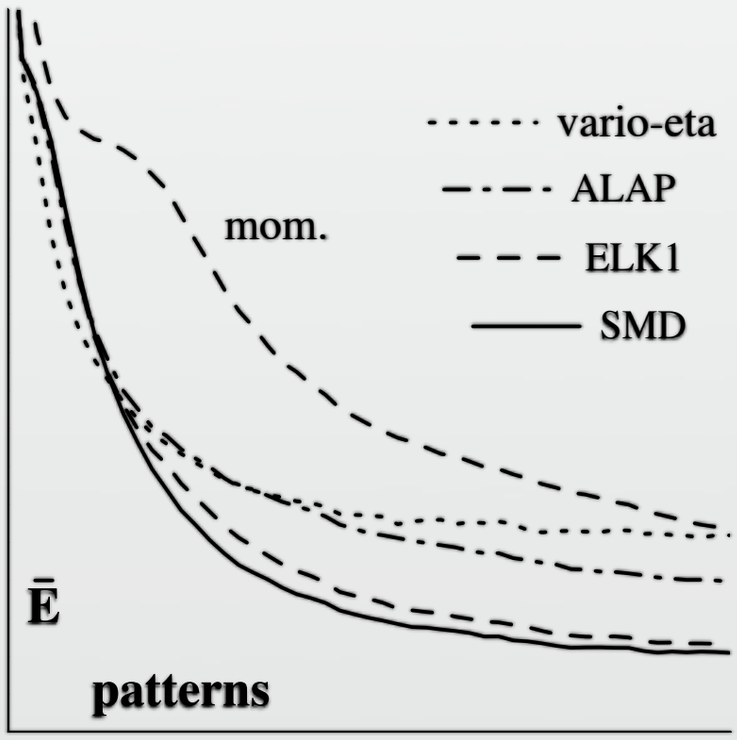


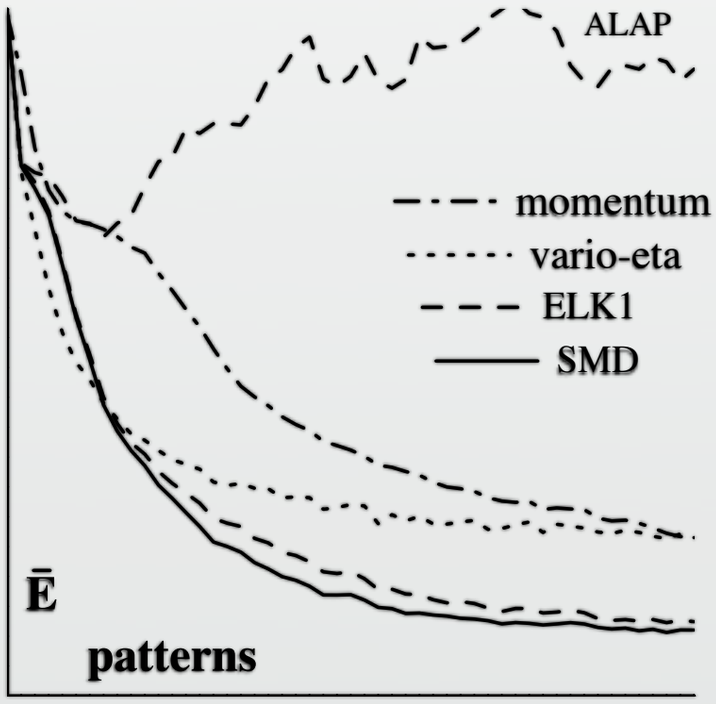
i.i.d. uniform

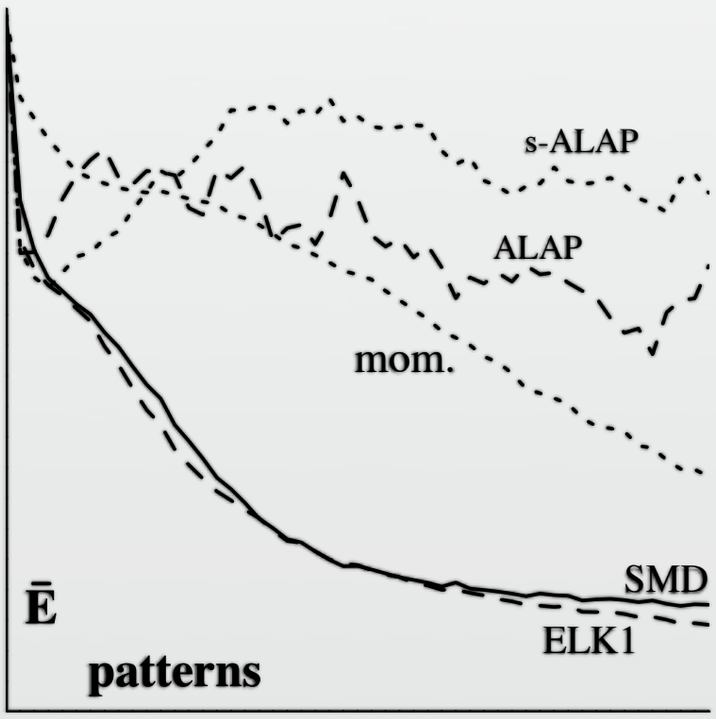


Sobol

Brownian







Comparison to CG

Conjugate Gradient

deterministic (1000 pts)

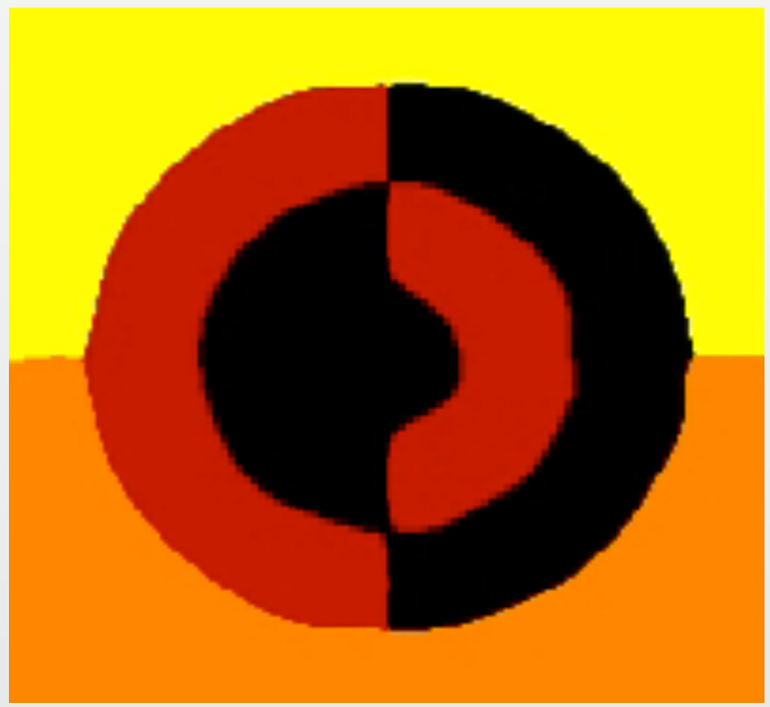
overfits

stochastic (1000 pts/iteration)

diverges

SMD

stochastic (5 pts/iteration)



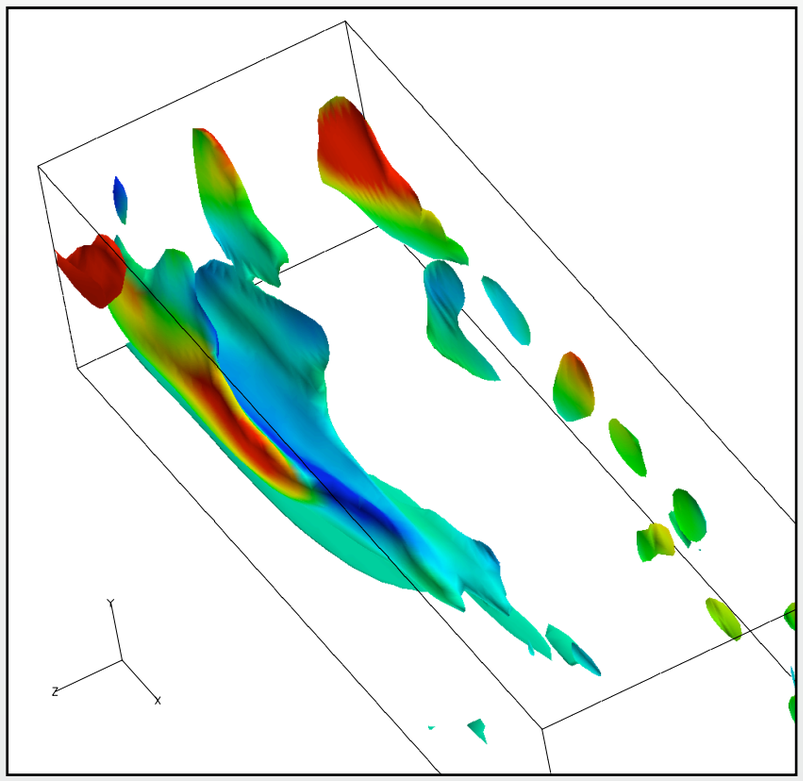
converges

SMD

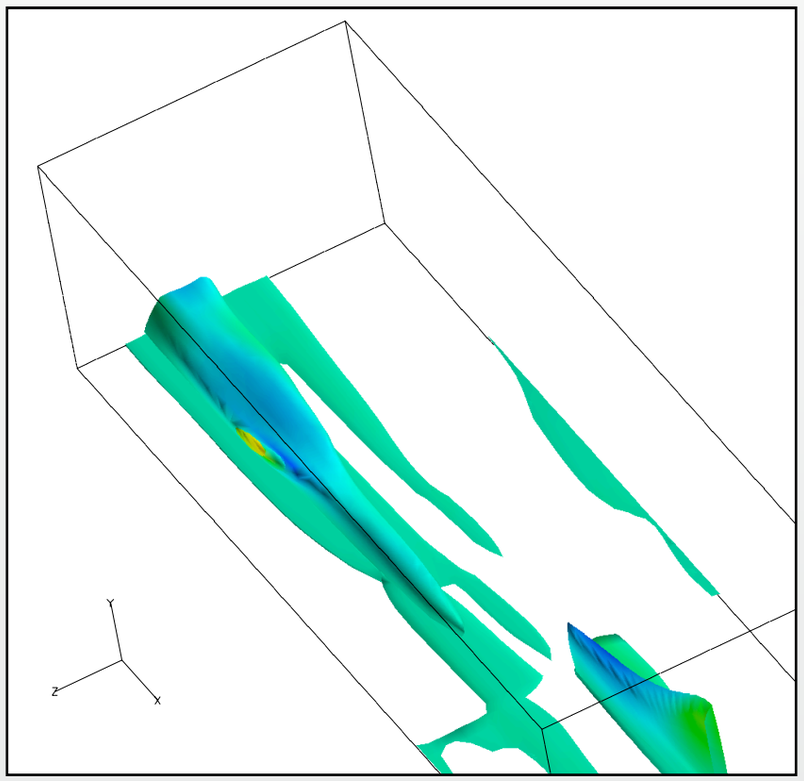
Applications

Application: Turbulent Flow

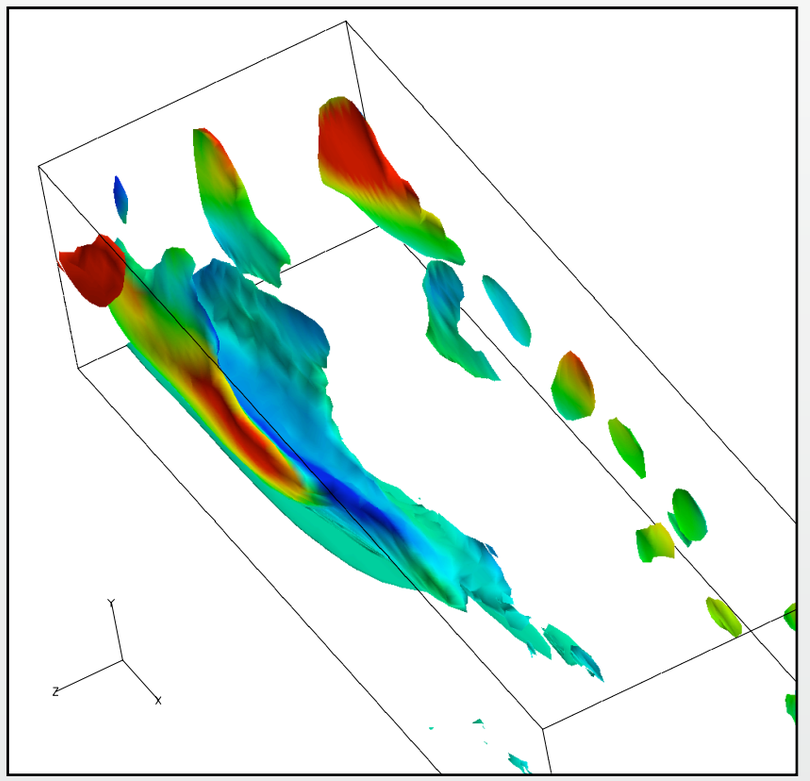
(PhD thesis of M. Milano, Inst. of Computational Science, ETH Zürich)



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

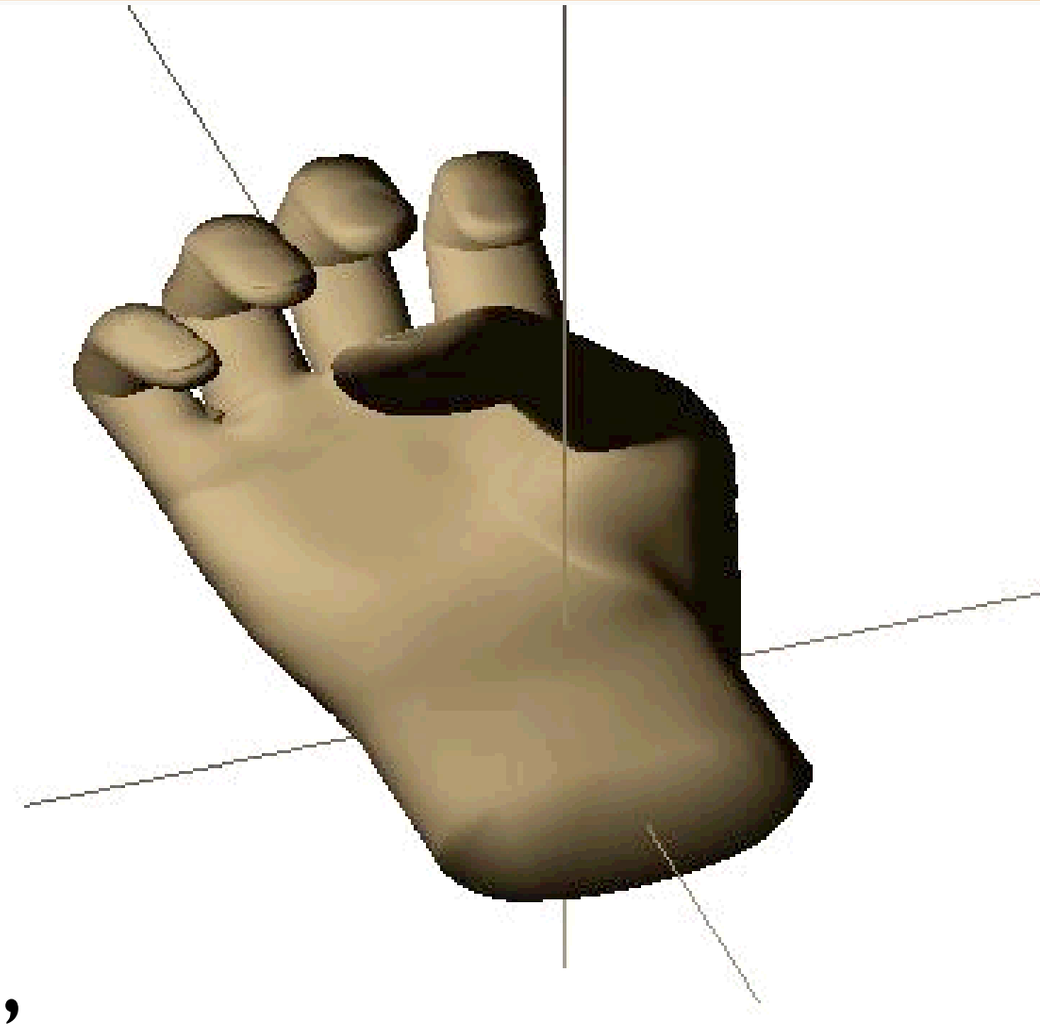
(114 sec/frame)

Stochastic Meta-Descent

(3 sec/frame)

Hand Tracking @ NICTA

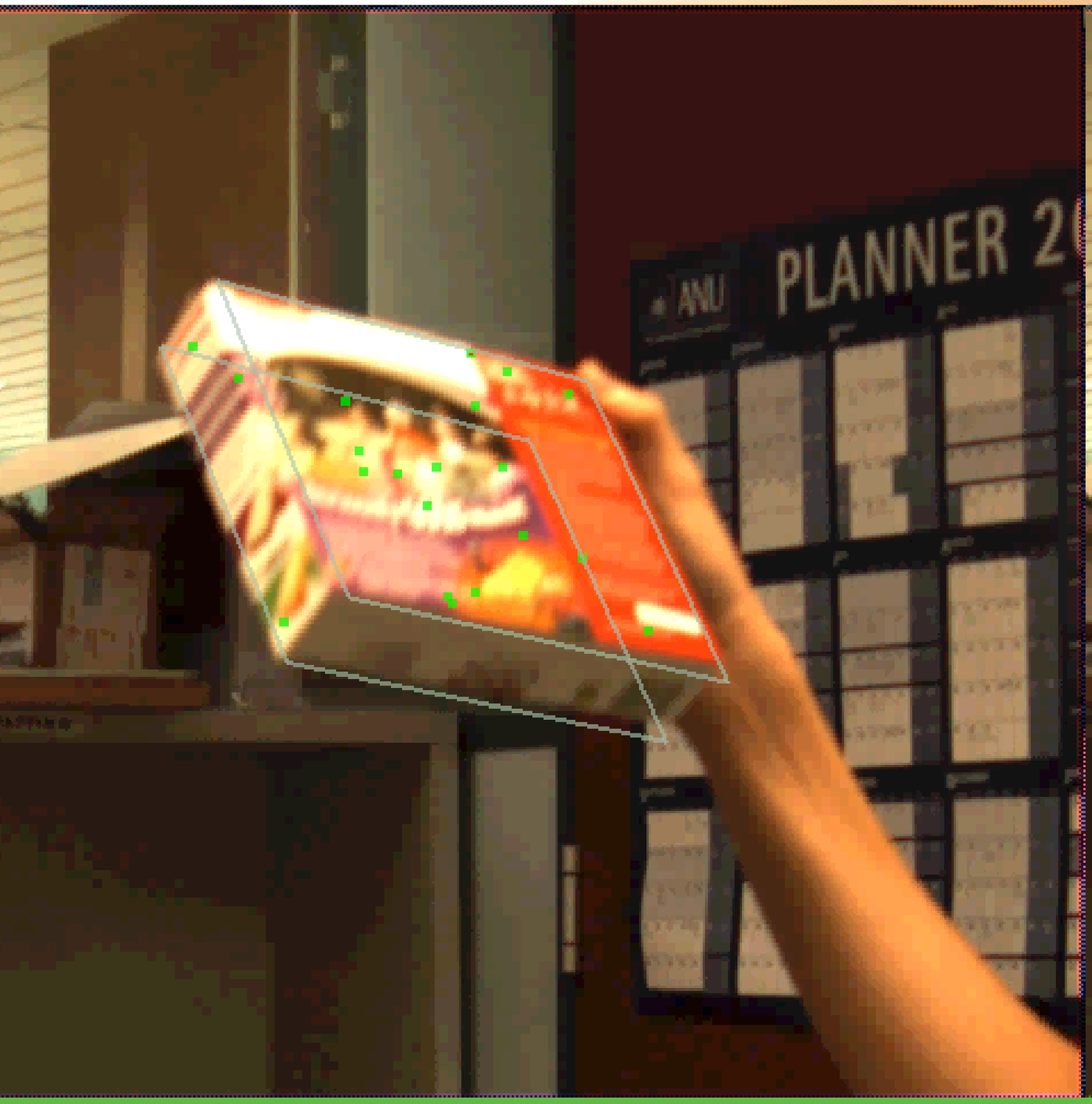
with Desmond Chik (PhD)
and Jochen Trumpf (SEACS)

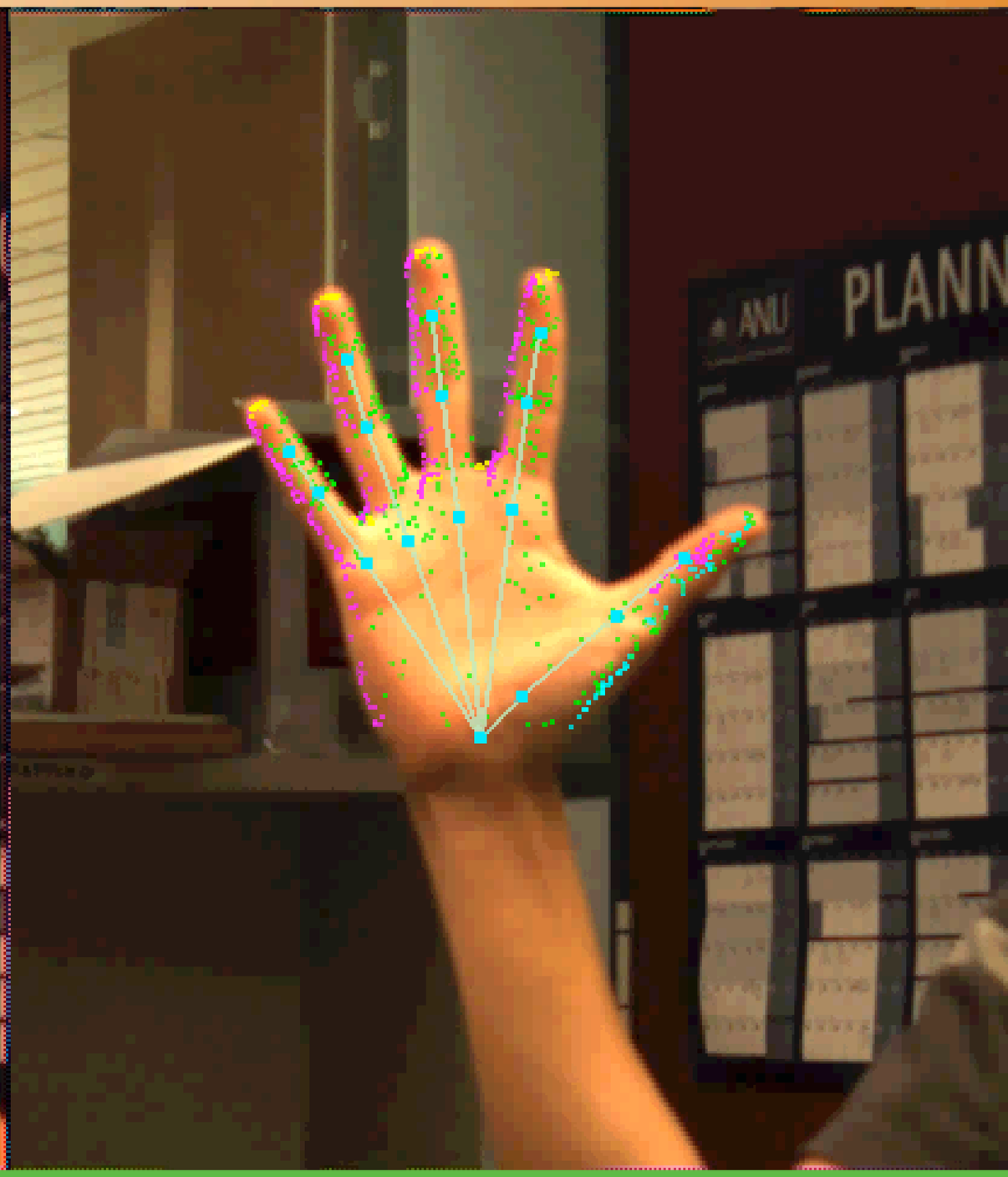


- detailed hand model (26 dof, ~10k vertices, skin blending, ...)
- randomly sample few points on model surface, project to image
- compare camera image(s) there, use resulting stochastic gradient to adjust model
- radical: completely marker- and feature-free tracking

NATIONAL ICT AUSTRALIA

Hand (et al.) Tracking

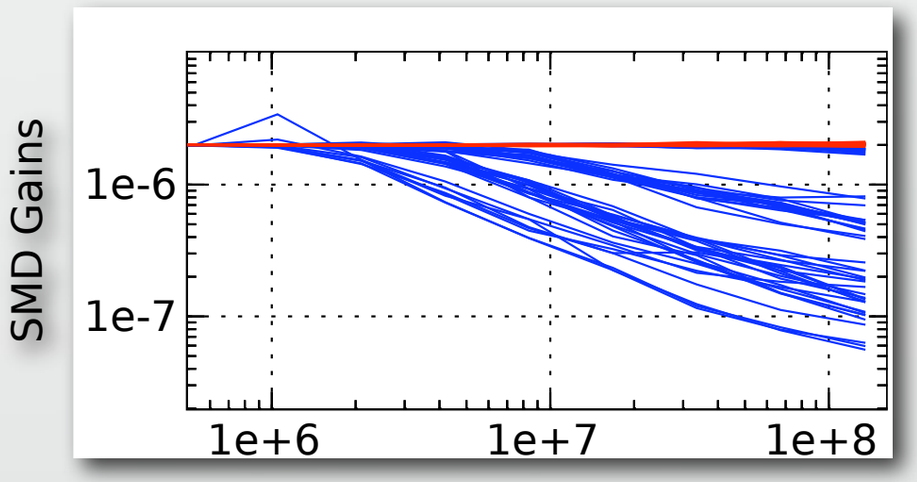


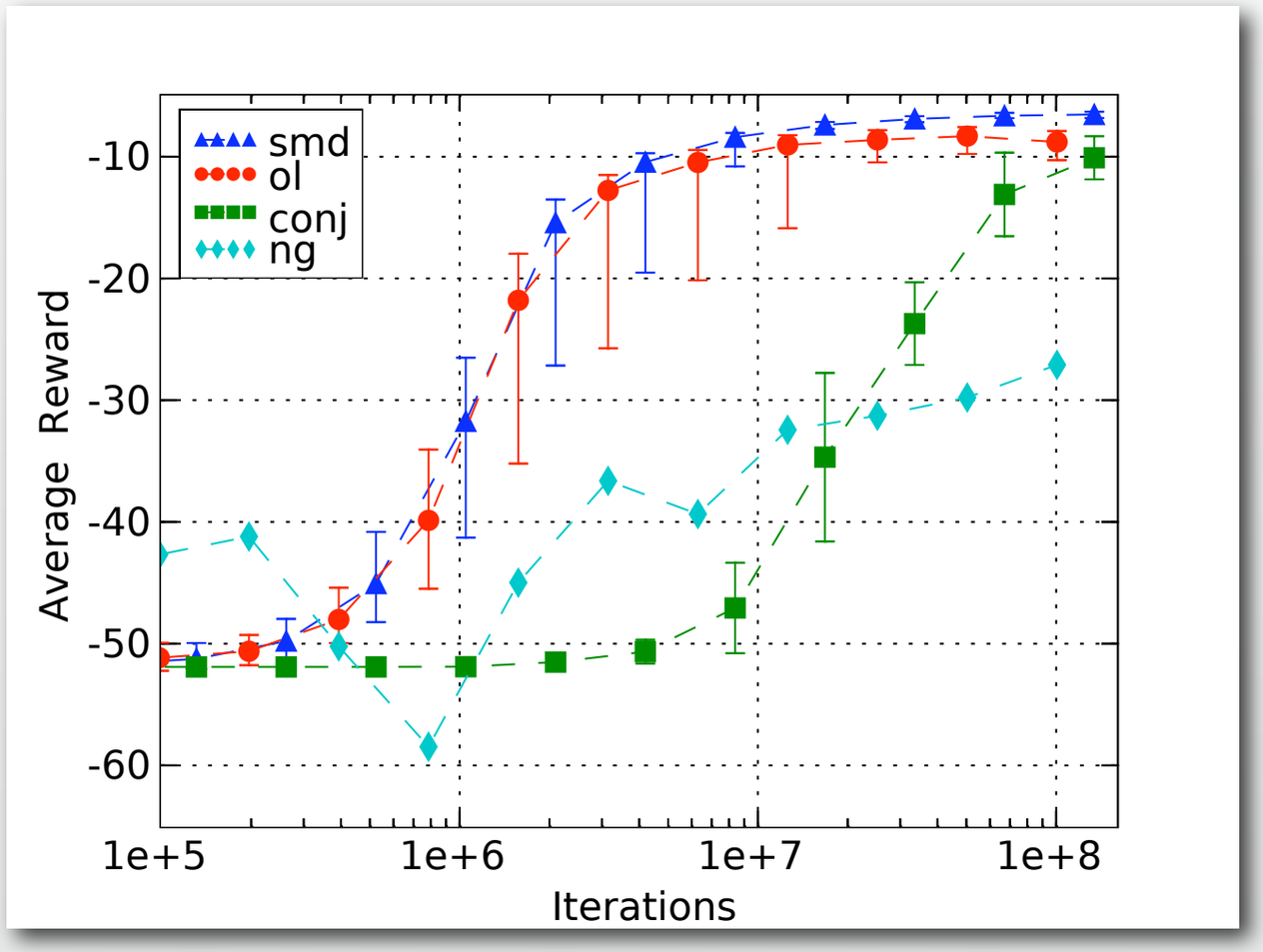


SMD & Policy Gradient RL

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

- SMD accelerates PG-RL
- complex interaction with temporal task structure
- had a poster at NIPS'05

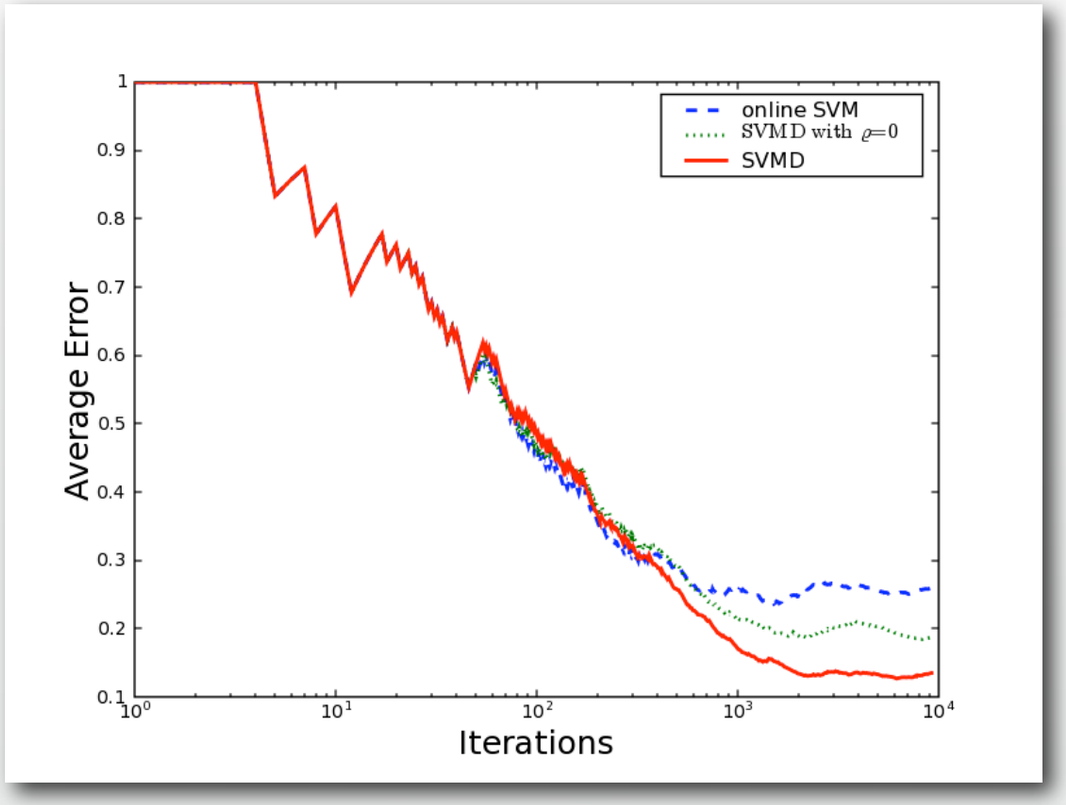




SMD for Online SVM

Online SVM aka NORMA (Kivinen, Smola, Williamson 2004):

- online kernel method
- stochastic gradient in expansion coefficients
- employs scalar gain η

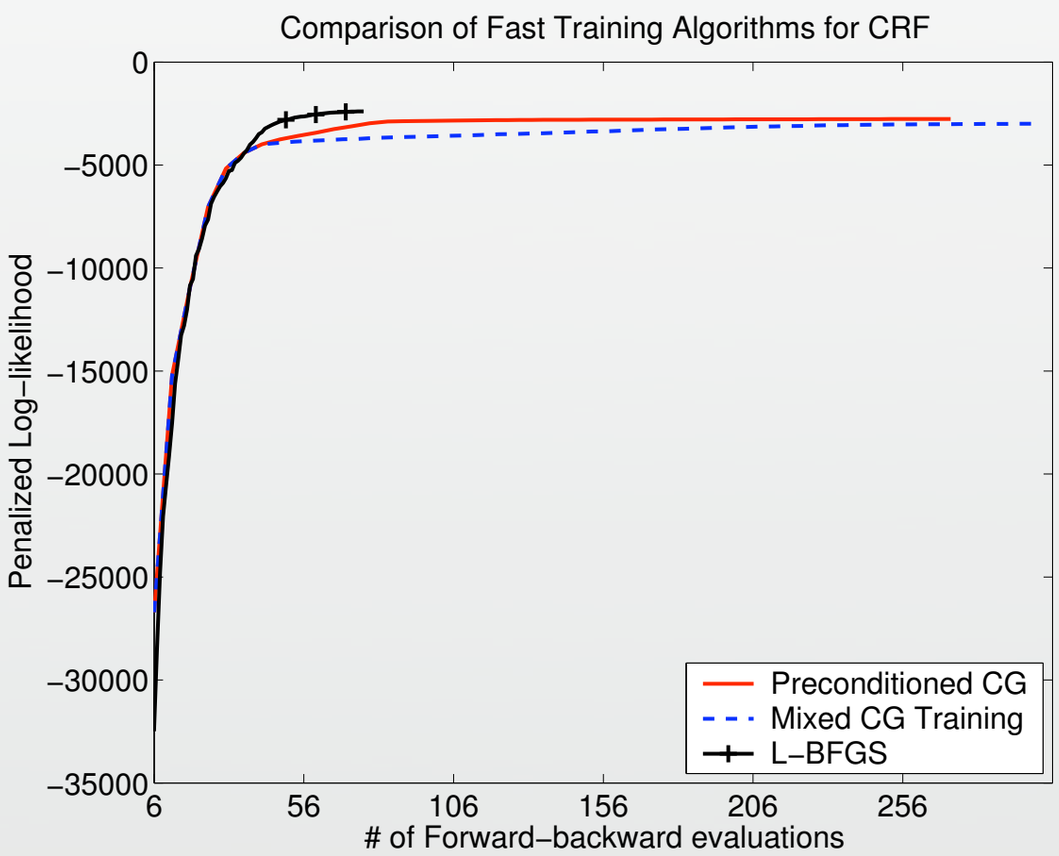


Applied SMD (with Vishy):

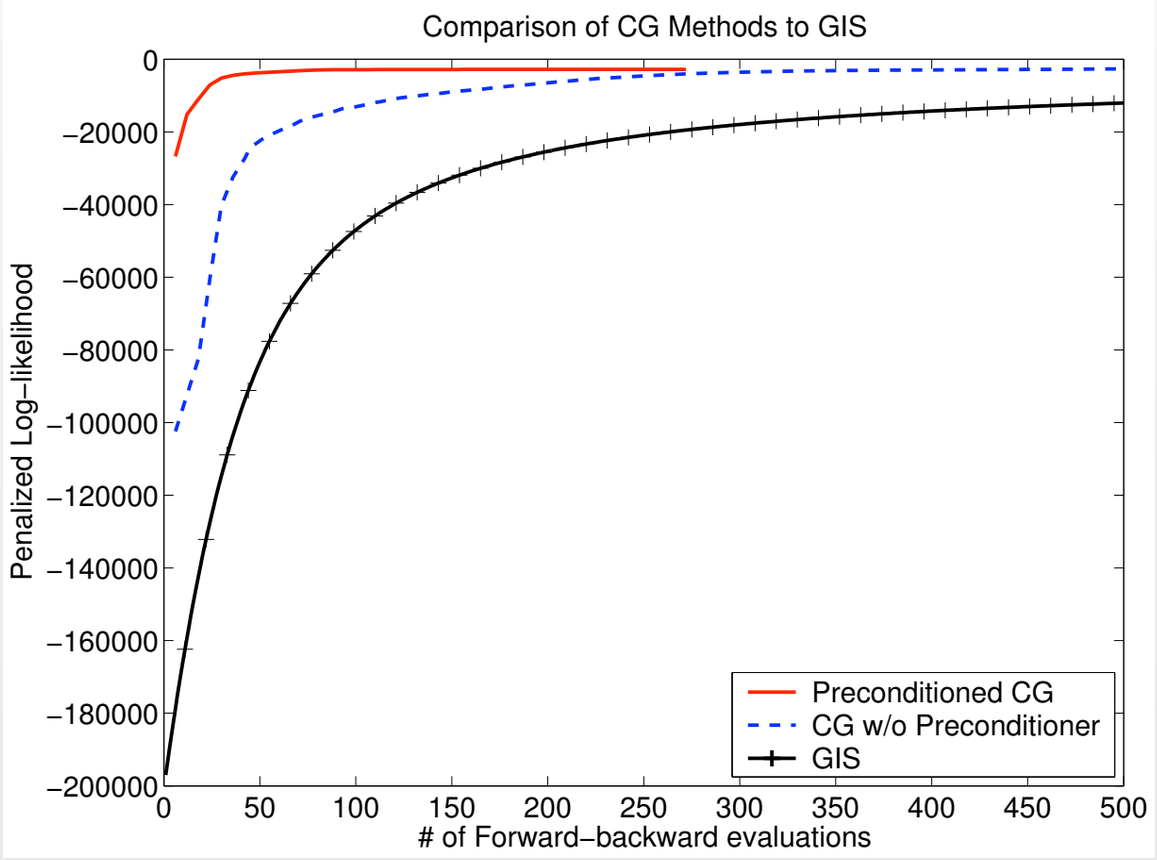
- **v** is function in RKHS
- <**g**,**v**> can be maintained incrementally in O(n)
- presented at NIPS'05 workshop, submitted to JMLR

Training CRFs (CoNLL-2000)

Sha, F. & Pereira, F. (2003). Shallow parsing with conditional random fields. In Proceedings of HLT-NAACL, 213–220. Association for Computational Linguistics.

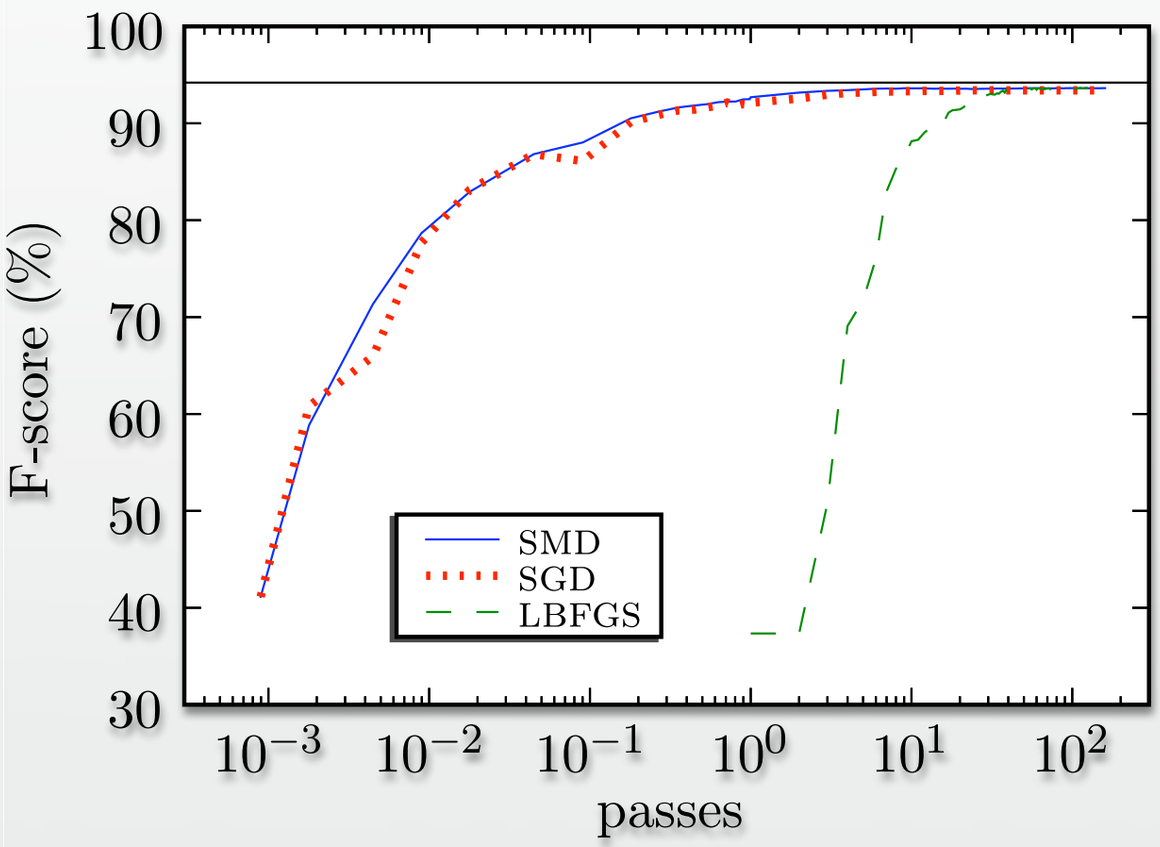


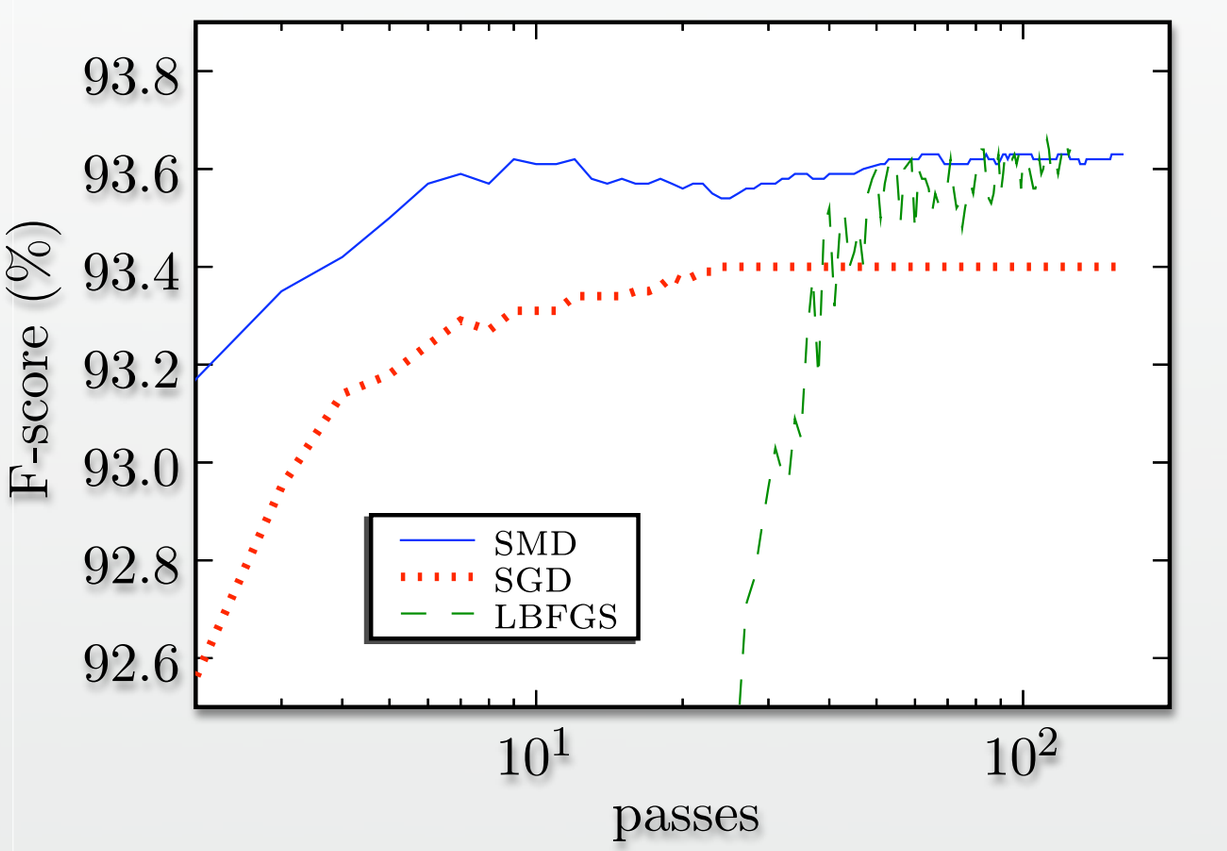
(a) $\mathcal{L}'_{\boldsymbol{\lambda}}$: CG (precond., mixed), L-BFGS



(b) $\mathcal{L}'_{\boldsymbol{\lambda}}$: CG (precond., plain), GIS

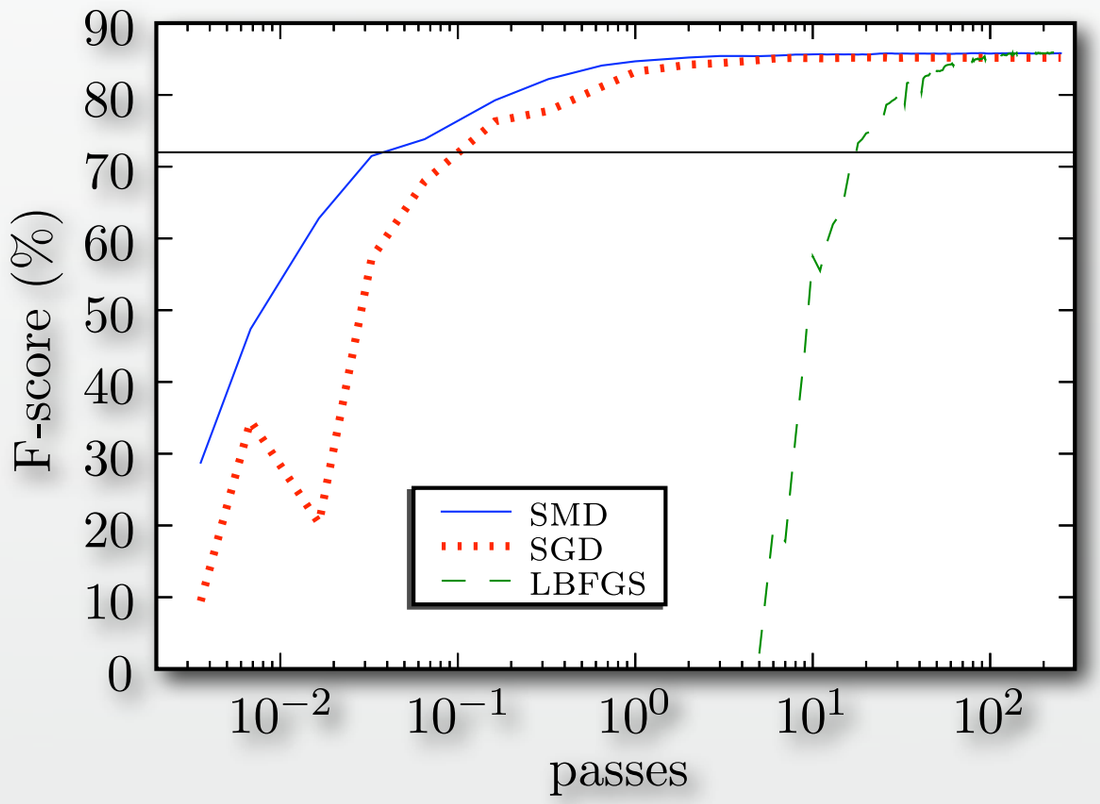
Training CRFs with SMD

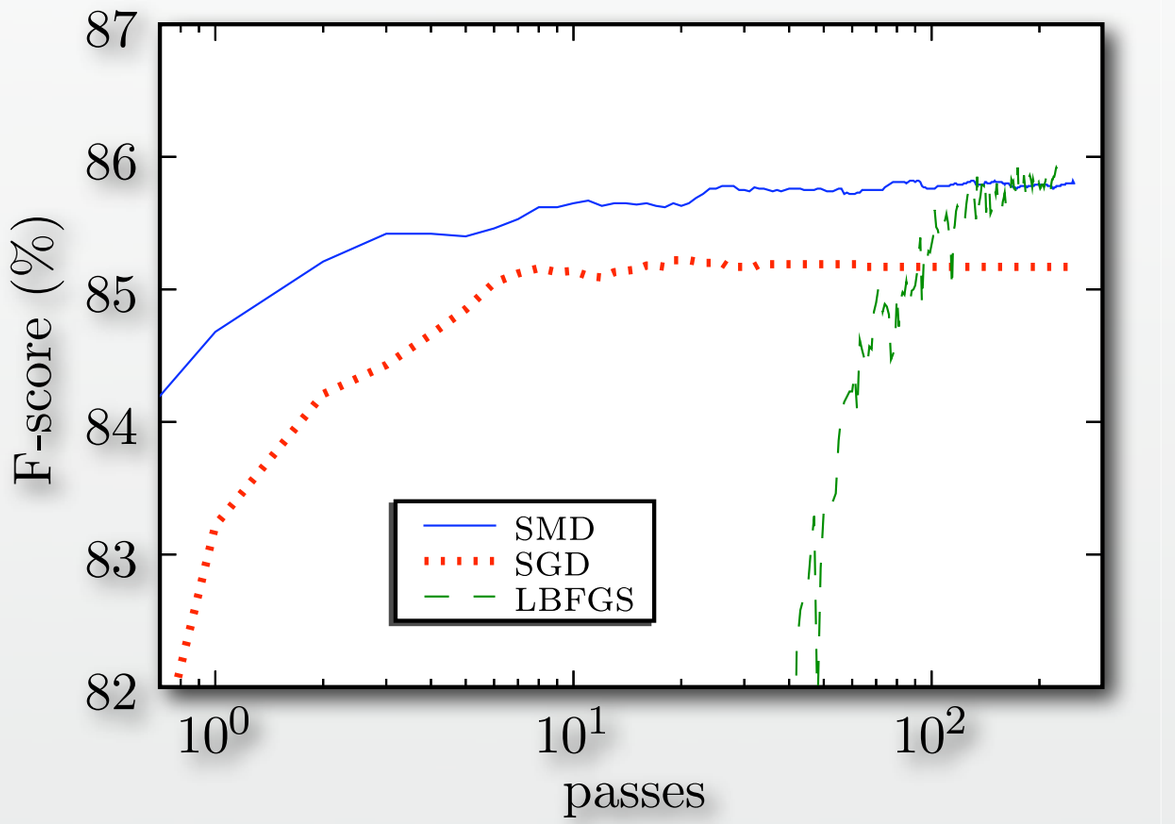




- 1-D CRF chain for ConNLL-2000 Base NP chunking
- (predictable!) huge speed-up for online learning

Training CRFs with SMD

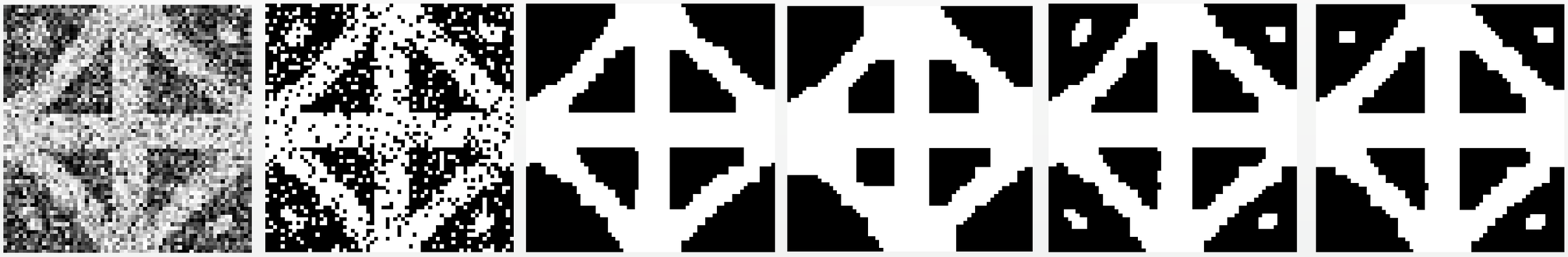




- BioNLP/NLPBA-2004 named-entity recognition task
- 2-D CRF lattices for vision (M. Schmidt & K. Murphy, UBC)

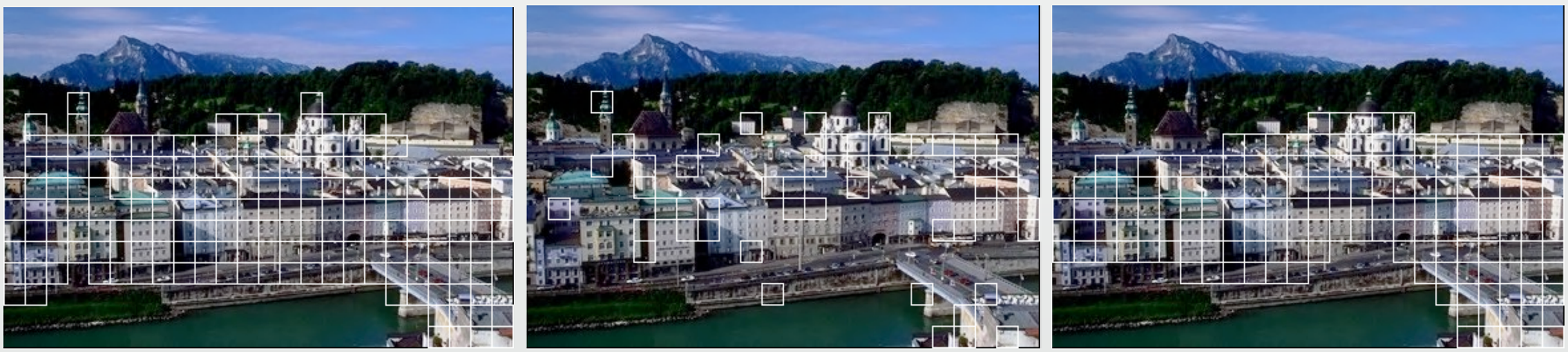
  work well with loopy BP; more robust to overfitting

SMD on 2-D CRF Lattices



original log. regr. BFGS SGD SMD SMD (PL)

ground truth logistic regression SMD (loopy BP)



Summary and Outlook

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** 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