Adaptive Subgradient Methods for Online Learning and Stochastic Optimization

John C. Duchi[1,2] Elad Hazan[3] Yoram Singer[2]

[1]University of California, Berkeley

[2]Google Research

[3]Technion

International Symposium on Mathematical Programming 2012

Setting: Online Convex Optimization

Online learning task—repeat:

- Learner plays point x_t
- Receive function f_t
- Suffer loss
  $f_t(x_t) + \varphi(x_t)$

Setting: Online Convex Optimization

Online learning task—repeat:

- Learner plays point x_t
- Receive function f_t
- Suffer loss $f_t(x_t) + \varphi(x_t)$

- Parameter vector for features
- Receive label y_t, features ϕ_t
- Suffer regularized logistic loss $\log\left[1 + \exp(-y_t \langle \phi_t, x_t \rangle)\right] + \lambda \left\|x_t\right\|_1$

Setting: Online Convex Optimization

Online learning task—repeat:

- Learner plays point x_t
- Receive function f_t
- Suffer loss $f_t(x_t) + \varphi(x_t)$

- Parameter vector for features
- Receive label y_t, features ϕ_t
- Suffer regularized logistic loss $\log\left[1 + \exp(-y_t \langle \phi_t, x_t \rangle)\right] + \lambda \|x_t\|_1$

Goal: Attain small regret

$$\sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - \inf_{x \in \mathcal{X}} \left[\sum_{t=1}^{T} f_t(x) + \varphi(x) \right]$$

Motivation

Text data:

`The most unsung birthday in American business and technological history this year may be the 50th anniversary of the Xerox 914 photocopier.`[a]

[a] *The Atlantic*, July/August 2010.

Motivation

Text data:

`The most unsung birthday in American business and technological history this year may be the 50th anniversary of the Xerox 914 photocopier.`[a]

[a] *The Atlantic*, July/August 2010.

High-dimensional image features

Motivation

Text data:

`The most unsung birthday in American business and technological history this year may be the 50th anniversary of the Xerox 914 photocopier.`[a]

[a] *The Atlantic*, July/August 2010.

High-dimensional image features

Other motivation: selecting advertisements in online advertising, document ranking, problems with parameterizations of many magnitudes...

Goal?

Flipping around the usual sparsity game

$$\min_x . \ \|Ax - b\|, \quad A = [a_1 \ a_2 \ \cdots \ a_n]^\top \in \mathbb{R}^{n \times d}$$

Usually in sparsity-focused depend on

$$\underbrace{\|a_i\|_\infty}_{\text{dense}} \cdot \underbrace{\|x\|_1}_{\text{sparse}}$$

Goal?

Flipping around the usual sparsity game

$$\min_x . \ \|Ax - b\|, \quad A = [a_1 \ a_2 \ \cdots \ a_n]^\top \in \mathbb{R}^{n \times d}$$

Usually in sparsity-focused depend on

$$\underbrace{\|a_i\|_\infty}_{\text{dense}} \cdot \underbrace{\|x\|_1}_{\text{sparse}}$$

What we would like:

$$\underbrace{\|a_i\|_1}_{\text{sparse}} \cdot \underbrace{\|x\|_\infty}_{\text{dense}}$$

Goal?

Flipping around the usual sparsity game

$$\min_x . \ \|Ax - b\|, \quad A = [a_1 \ a_2 \ \cdots \ a_n]^\top \in \mathbb{R}^{n \times d}$$

Usually in sparsity-focused depend on

$$\underbrace{\|a_i\|_\infty}_{\text{dense}} \cdot \underbrace{\|x\|_1}_{\text{sparse}}$$

What we would like:

$$\underbrace{\|a_i\|_1}_{\text{sparse}} \cdot \underbrace{\|x\|_\infty}_{\text{dense}}$$

(In general, impossible)

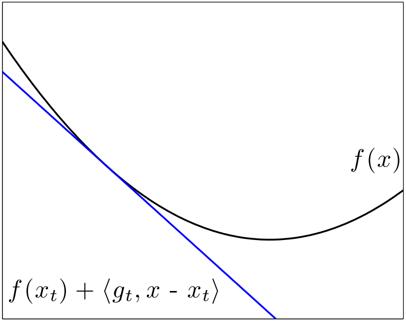
Approaches: Gradient Descent and Dual Averaging

Let $g_t \in \partial f_t(x_t)$:

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta_t \langle g_t, x \rangle \right\}$$

or

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{\eta_t}{t} \sum_{\tau=1}^{t} \langle g_\tau, x \rangle + \frac{1}{2t} \|x\|^2 \right\}$$



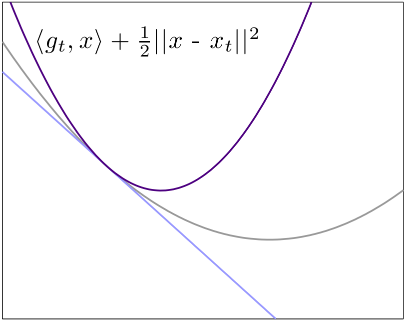
Approaches: Gradient Descent and Dual Averaging

Let $g_t \in \partial f_t(x_t)$:

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta_t \langle g_t, x \rangle \right\}$$

or

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{\eta_t}{t} \sum_{\tau=1}^{t} \langle g_\tau, x \rangle + \frac{1}{2t} \|x\|^2 \right\}$$



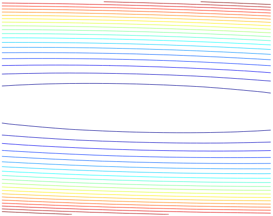
What is the problem?

- Gradient steps treat all features as equal

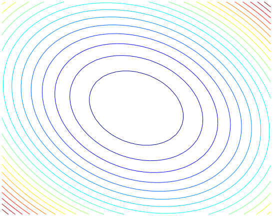
- They are not!

Adapting to Geometry of Space

Why adapt to geometry?

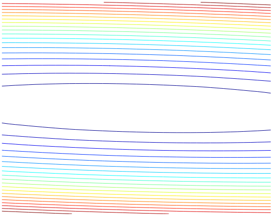


Hard

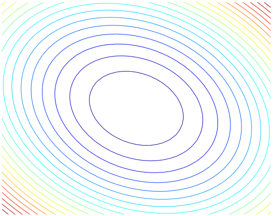


Nice

Why adapt to geometry?



Hard

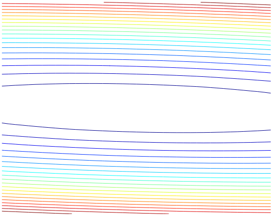


Nice

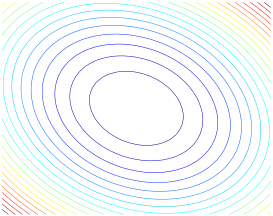
y_t	$\phi_{t,1}$	$\phi_{t,2}$	$\phi_{t,3}$
1	1	0	0
-1	.5	0	1
1	-.5	1	0
-1	0	0	0
1	.5	0	0
-1	1	0	0
1	-1	1	0
-1	-.5	0	1

1. Frequent, irrelevant
2. Infrequent, predictive
3. Infrequent, predictive

Why adapt to geometry?



Hard



Nice

y_t	$\phi_{t,1}$	$\phi_{t,2}$	$\phi_{t,3}$
1	1	0	0
-1	.5	0	1
1	-.5	1	0
-1	0	0	0
1	.5	0	0
-1	1	0	0
1	-1	1	0
-1	-.5	0	1

1. Frequent, irrelevant
2. Infrequent, predictive
3. Infrequent, predictive

Adapting to Geometry of the Space

- Receive $g_t \in \partial f_t(x_t)$
- Earlier:

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|^2 + \eta \langle g_t, x \rangle \right\}$$

Adapting to Geometry of the Space

- Receive $g_t \in \partial f_t(x_t)$
- Earlier:

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \left\| x - x_t \right\|^2 + \eta \left\langle g_t, x \right\rangle \right\}$$

- Now: let $\|x\|_A^2 = \langle x, Ax \rangle$ for $A \succeq 0$. Use

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \left\| x - x_t \right\|_A^2 + \eta \left\langle g_t, x \right\rangle \right\}$$

Regret Bounds

What does adaptation buy?

- Standard regret bound:

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \le \frac{1}{2\eta} \|x_1 - x^*\|_2^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_2^2$$

Regret Bounds

What does adaptation buy?

- Standard regret bound:

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \frac{1}{2\eta} \|x_1 - x^*\|_2^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_2^2$$

- Regret bound with matrix:

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \frac{1}{2\eta} \|x_1 - x^*\|_A^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_{A^{-1}}^2$$

Meta Learning Problem

- Have regret:

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \frac{1}{\eta} \|x_1 - x^*\|_A^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|_{A^{-1}}^2$$

- What happens if we minimize A in hindsight?

$$\min_A \sum_{t=1}^{T} \left\langle g_t, A^{-1} g_t \right\rangle \quad \text{subject to} \quad A \succeq 0, \mathrm{tr}(A) \leq C$$

Meta Learning Problem

- What happens if we minimize A in hindsight?

$$\min_A \sum_{t=1}^{T} \left\langle g_t, A^{-1} g_t \right\rangle \quad \text{subject to} \quad A \succeq 0, \operatorname{tr}(A) \leq C$$

Meta Learning Problem

- What happens if we minimize A in hindsight?

$$\min_{A} \sum_{t=1}^{T} \left\langle g_t, A^{-1} g_t \right\rangle \quad \text{subject to} \quad A \succeq 0, \mathrm{tr}(A) \leq C$$

- Solution is of form

$$A = c \, \mathrm{diag}\left(\sum_{t=1}^{T} g_t g_t^\top \right)^{\frac{1}{2}} \qquad A = c\left(\sum_{t=1}^{T} g_t g_t^\top \right)^{\frac{1}{2}}$$

(diagonal) (full)

(where c chosen to satisfy tr constraint)

Meta Learning Problem

- What happens if we minimize A in hindsight?

$$\min_A \sum_{t=1}^{T} \left\langle g_t, A^{-1} g_t \right\rangle \quad \text{subject to} \quad A \succeq 0, \mathrm{tr}(A) \leq C$$

- Solution is of form

$$A = c \, \mathrm{diag}\left(\sum_{t=1}^{T} g_t g_t^\top\right)^{\frac{1}{2}} \qquad A = c\left(\sum_{t=1}^{T} g_t g_t^\top\right)^{\frac{1}{2}}$$

(diagonal) (full)

(where c chosen to satisfy tr constraint)

- Let $g_{1:t,j}$ be vector of jth gradient component. Optimal:

$$A_{j,j} \propto \|g_{1:T}, j\|_2$$

Low regret to the best A

- Let $g_{1:t,j}$ be vector of jth gradient component. At time t, use

$$s_t = \left[\|g_{1:t,j}\|_2\right]_{j=1}^d \quad \text{and} \quad A_t = \mathrm{diag}(s_t)$$

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|_{A_t}^2 + \eta \langle g_t, x \rangle \right\}$$

Low regret to the best A

- Let $g_{1:t,j}$ be vector of jth gradient component. At time t, use

$$s_t = \left[\|g_{1:t,j}\|_2\right]_{j=1}^{d} \quad \text{and} \quad A_t = \operatorname{diag}(s_t)$$

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|_{A_t}^2 + \eta \langle g_t, x \rangle \right\}$$

- Example:

y_t	$g_{t,1}$	$g_{t,2}$	$g_{t,3}$
1	1	0	0
-1	.5	0	1
1	-.5	1	0
-1	0	0	0
1	1	1	0
-1	1	0	0
	$s_1 = \sqrt{3.5}$	$s_2 = \sqrt{2}$	$s_3 = 1$

Final Convergence Guarantee

Algorithm: at time t, set

$$s_t = \left[\|g_{1:t,j}\|_2\right]_{j=1}^d \quad \text{and} \quad A_t = \mathrm{diag}(s_t)$$

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|_{A_t}^2 + \eta \langle g_t, x \rangle \right\}$$

Define radius

$$R_\infty := \max_t \|x_t - x^*\|_\infty \le \sup_{x \in \mathcal{X}} \|x - x^*\|_\infty .$$

Final Convergence Guarantee

Algorithm: at time t, set

$$s_t = \left[\|g_{1:t,j}\|_2\right]_{j=1}^d \quad \text{and} \quad A_t = \mathrm{diag}(s_t)$$

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|_{A_t}^2 + \eta \langle g_t, x \rangle \right\}$$

Define radius

$$R_\infty := \max_t \|x_t - x^*\|_\infty \leq \sup_{x \in \mathcal{X}} \|x - x^*\|_\infty .$$

Theorem

The final regret bound of AdaGrad:

$$\sum_{t=1}^T f_t(x_t) - f_t(x^*) \leq 2R_\infty \sum_{i=1}^d \|g_{1:T,j}\|_2 .$$

Understanding the convergence guarantees I

- Stochastic convex optimization:

$$f(x) := \mathbb{E}_P[f(x;\xi)]$$

Understanding the convergence guarantees I

- Stochastic convex optimization:

$$f(x) := \mathbb{E}_P[f(x;\xi)]$$

Sample ξ_t according to P, define $f_t(x) := f(x;\xi_t)$. Then

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f(x^*) \le \frac{2R_\infty}{T}\sum_{i=1}^{d}\mathbb{E}\left[\left\|g_{1:T,j}\right\|_2\right]$$

Understanding the convergence guarantees II

Support vector machine example: define

$$f(x; \xi) = [1 - \langle x, \xi \rangle]_+ , \quad \text{where } \xi \in \{-1, 0, 1\}^d$$

Understanding the convergence guarantees II

Support vector machine example: define

$$f(x;\xi) = [1 - \langle x, \xi \rangle]_+ , \qquad \text{where } \xi \in \{-1, 0, 1\}^d$$

- If $\xi_j \neq 0$ with probability $\propto j^{-\alpha}$ for $\alpha > 1$

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f(x^*) = \mathcal{O}\left(\frac{\|x^*\|_\infty}{\sqrt{T}} \cdot \max\left\{\log d, d^{1-\alpha/2}\right\}\right)$$

Understanding the convergence guarantees II

Support vector machine example: define

$$f(x;\xi) = [1 - \langle x, \xi \rangle]_+, \quad \text{where } \xi \in \{-1, 0, 1\}^d$$

- If $\xi_j \neq 0$ with probability $\propto j^{-\alpha}$ for $\alpha > 1$

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f(x^*) = \mathcal{O}\left(\frac{\|x^*\|_\infty}{\sqrt{T}} \cdot \max\left\{\log d, d^{1-\alpha/2}\right\}\right)$$

- Previously best-known method:

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f(x^*) = \mathcal{O}\left(\frac{\|x^*\|_\infty}{\sqrt{T}} \cdot \sqrt{d}\right).$$

Understanding the convergence guarantees III

Back to regret minimization

- Convergence almost as good as that of the best geometry matrix:

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*)$$

$$\leq 2\sqrt{d}\,\|x^*\|_\infty \sqrt{\inf_s \left\{ \sum_{t=1}^{T} \|g_t\|^2_{\mathrm{diag}(s)^{-1}} \;:\; s \succeq 0,\ \langle \mathbb{1}, s\rangle \leq d \right\}}$$

Understanding the convergence guarantees III

Back to regret minimization

- Convergence almost as good as that of the best geometry matrix:

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*)$$

$$\leq 2\sqrt{d}\,\|x^*\|_\infty \sqrt{\inf_s \left\{ \sum_{t=1}^{T} \|g_t\|^2_{\mathrm{diag}(s)^{-1}} \ : \ s \succeq 0, \ \langle \mathbb{1}, s\rangle \leq d \right\}}$$

- This (and other bounds) are minimax optimal

The AdaGrad Algorithms

Analysis applies to several algorithms

$$s_t = \left[\|g_{1:t,j}\|_2\right]_{j=1}^{d}, \qquad A_t = \mathrm{diag}(s_t)$$

The AdaGrad Algorithms

Analysis applies to several algorithms

$$s_t = \left[\|g_{1:t,j}\|_2\right]_{j=1}^d, \qquad A_t = \mathrm{diag}(s_t)$$

- Forward-backward splitting (Lions and Mercier 1979, Nesterov 2007, Duchi and Singer 2009, others)

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|_{A_t}^2 + \langle g_t, x \rangle + \varphi(x) \right\}$$

The AdaGrad Algorithms

Analysis applies to several algorithms

$$s_t = \left[\|g_{1:t,j}\|_2\right]_{j=1}^{d}, \qquad A_t = \mathrm{diag}(s_t)$$

- Forward-backward splitting (Lions and Mercier 1979, Nesterov 2007, Duchi and Singer 2009, others)

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{2} \|x - x_t\|_{A_t}^2 + \langle g_t, x \rangle + \varphi(x) \right\}$$

- Regularized Dual Averaging (Nesterov 2007, Xiao 2010)

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \frac{1}{t} \sum_{\tau=1}^{t} \langle g_\tau, x \rangle + \varphi(x) + \frac{1}{2t} \|x\|_{A_t}^2 \right\}$$

An Example and Experimental Results

- ℓ_1-regularization
- Text classification
- Image ranking
- Neural network learning

AdaGrad with composite updates

Recall more general problem:

$$\sum_{t=1}^{T} f(x_t) + \varphi(x_t) - \inf_{x^* \in \mathcal{X}} \left[\sum_{t=1}^{T} f(x) + \varphi(x) \right]$$

AdaGrad with composite updates

Recall more general problem:

$$\sum_{t=1}^{T} f(x_t) + \varphi(x_t) - \inf_{x^* \in \mathcal{X}} \left[\sum_{t=1}^{T} f(x) + \varphi(x) \right]$$

- Must solve updates of form

$$\operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle g, x \rangle + \varphi(x) + \frac{1}{2} \|x\|_A^2 \right\}$$

AdaGrad with composite updates

Recall more general problem:

$$\sum_{t=1}^{T} f(x_t) + \varphi(x_t) - \inf_{x^* \in \mathcal{X}} \left[\sum_{t=1}^{T} f(x) + \varphi(x) \right]$$

- Must solve updates of form

$$\operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle g, x \rangle + \varphi(x) + \frac{1}{2} \|x\|_A^2 \right\}$$

- Luckily, often *still simple*

AdaGrad with ℓ_1 regularization

Set $\bar{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau$. Need to solve

$$\min_x \; \langle \bar{g}_t, x\rangle + \lambda \left\|x\right\|_1 + \frac{1}{2t}\left\langle x, \mathrm{diag}(s_t)x\right\rangle$$

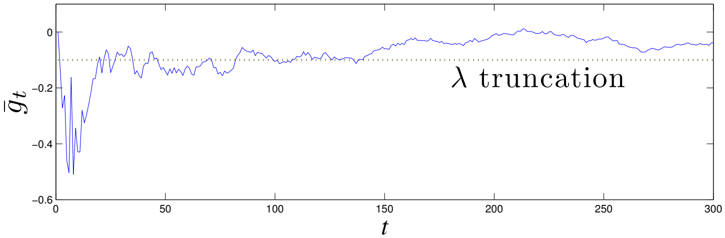
AdaGrad with ℓ_1 regularization

Set $\bar{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau$. Need to solve

$$\min_x \ \langle \bar{g}_t, x\rangle + \lambda \|x\|_1 + \frac{1}{2t}\langle x, \mathrm{diag}(s_t)x\rangle$$

- Coordinate-wise update yields sparsity and adaptivity:

$$x_{t+1,j} = \mathrm{sign}(-\bar{g}_{t,j})\frac{t}{\|g_{1:t,j}\|_2}\left[|\bar{g}_{t,j}| - \lambda\right]_+$$



Text Classification

Reuters RCV1 document classification task—$d = 2 \cdot 10^6$ features, approximately 4000 non-zero features per document

$$f_t(x) := [1 - \langle x, \xi_t \rangle]_+$$

where $\xi_t \in \{-1, 0, 1\}^d$ is data sample

[1] Crammer et al., 2006

[2] Crammer et al., 2009

Text Classification

Reuters RCV1 document classification task—$d = 2 \cdot 10^6$ features, approximately 4000 non-zero features per document

$$f_t(x) := [1 - \langle x, \xi_t \rangle]_+$$

where $\xi_t \in \{-1, 0, 1\}^d$ is data sample

	FOBOS	AdaGrad	PA[1]	AROW[2]
Ecomonics	.058 (.194)	**.044 (.086)**	.059	.049
Corporate	.111 (.226)	**.053 (.105)**	.107	.061
Government	.056 (.183)	**.040 (.080)**	.066	.044
Medicine	.056 (.146)	**.035 (.063)**	.053	.039

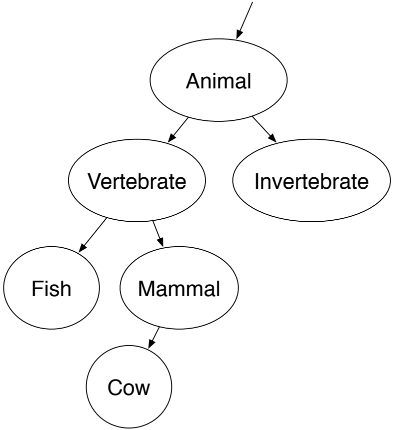
Test set classification error rate
(sparsity of final predictor in parenthesis)

[1] Crammer et al., 2006

[2] Crammer et al., 2009

Image Ranking

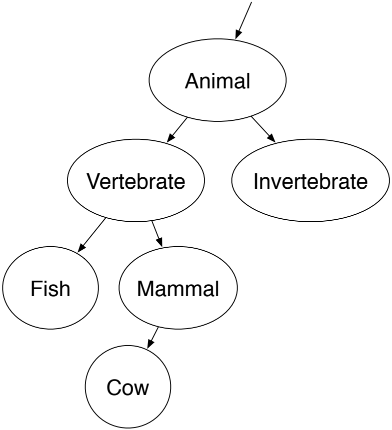
ImageNet (Deng et al., 2009), large-scale hierarchical image database



Train 15,000 rankers/classifiers to rank images for *each* noun (as in Grangier and Bengio, 2008)

Image Ranking

ImageNet (Deng et al., 2009), large-scale hierarchical image database



Train 15,000 rankers/classifiers to rank images for *each* noun (as in Grangier and Bengio, 2008)

Data
$\xi = (z^1, z^2) \in \{0,1\}^d \times \{0,1\}^d$ is pair of images

$$f(x; z^1, z^2) = \left[1 - \left\langle x, z^1 - z^2 \right\rangle\right]_+$$

Image Ranking Results

Precision at k: proportion of examples in top k that belong to category. Average precision is average placement of all positive examples.

Algorithm	Avg. Prec.	P@1	P@5	P@10	Nonzero
AdaGrad	**0.6022**	0.8502	0.8130	0.7811	**0.7267**
AROW	0.5813	**0.8597**	**0.8165**	**0.7816**	1.0000
PA	0.5581	0.8455	0.7957	0.7576	1.0000
Fobos	0.5042	0.7496	0.6950	0.6545	0.8996

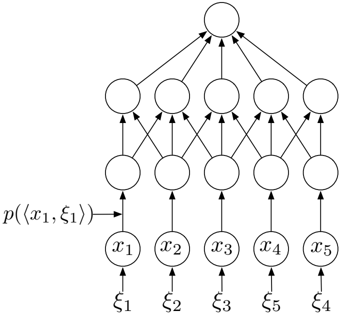
Neural Network Learning

Wildly non-convex problem:

$$f(x;\xi) = \log\left(1 + \exp\left(\langle [p(\langle x_1, \xi_1\rangle) \ \cdots \ p(\langle x_k, \xi_k\rangle)], \xi_0\rangle\right)\right)$$

where

$$p(\alpha) = \frac{1}{1+\exp(\alpha)}$$



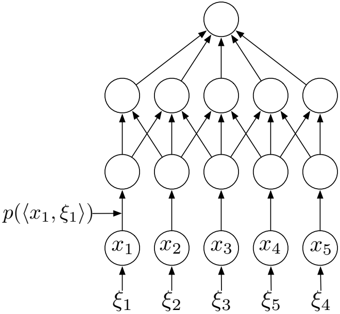
Neural Network Learning

Wildly non-convex problem:

$$f(x;\xi) = \log\left(1 + \exp\left(\langle [p(\langle x_1, \xi_1\rangle) \ \cdots \ p(\langle x_k, \xi_k\rangle)], \xi_0\rangle\right)\right)$$

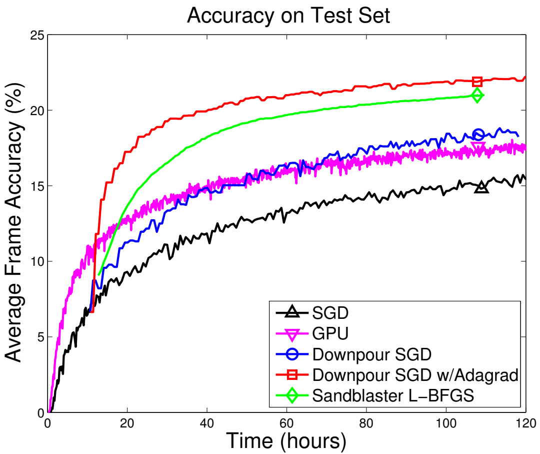
where

$$p(\alpha) = \frac{1}{1 + \exp(\alpha)}$$



Idea: Use stochastic gradient methods to solve it anyway

Neural Network Learning



(Dean et al. 2012)

Distributed, $d = 1.7 \cdot 10^9$ parameters. SGD and AdaGrad use 80 machines (1000 cores), L-BFGS uses 800 (10000 cores)

Conclusions and Discussion

- Family of algorithms that adapt to geometry of data
- Extendable to full matrix case to handle feature correlation
- Can derive many efficient algorithms for high-dimensional problems, especially with sparsity

Conclusions and Discussion

- Family of algorithms that adapt to geometry of data
- Extendable to full matrix case to handle feature correlation
- Can derive many efficient algorithms for high-dimensional problems, especially with sparsity
- Future: Efficient full-matrix adaptivity, other types of adaptation

Thanks!

OGD Sketch: "Almost" Contraction

- Have $g_t \in \partial f_t(x_t)$ (ignore φ, $\mathcal{X}$ for simplicity)
- Before: $x_{t+1} = x_t - \eta g_t$

$$\frac{1}{2}\left\|x_{t+1} - x^*\right\|_2^2 \le \frac{1}{2}\left\|x_t - x^*\right\|_2^2 + \eta\left(f_t(x^*) - f_t(x_t)\right) + \frac{\eta^2}{2}\left\|g_t\right\|_2^2$$

OGD Sketch: "Almost" Contraction

- Have $g_t \in \partial f_t(x_t)$ (ignore φ, $\mathcal{X}$ for simplicity)
- Before: $x_{t+1} = x_t - \eta g_t$

$$\frac{1}{2}\|x_{t+1} - x^*\|_2^2 \le \frac{1}{2}\|x_t - x^*\|_2^2 + \eta\left(f_t(x^*) - f_t(x_t)\right) + \frac{\eta^2}{2}\|g_t\|_2^2$$

- Now: $x_{t+1} = x_t - \eta A^{-1} g_t$

$$\begin{aligned}
&\frac{1}{2}\|x_{t+1} - x^*\|_A^2 \\
&= \frac{1}{2}\|x_t - x^*\|_A^2 + \eta\langle g_t, x^* - x_t\rangle + \frac{\eta^2}{2}\|g_t\|_{A^{-1}}^2 \\
&\le \frac{1}{2}\|x_t - x^*\|_A^2 + \eta\left(f_t(x^*) - f_t(x_t)\right) + \frac{\eta^2}{2}\underset{\substack{\uparrow \\ \text{dual norm to } \|\cdot\|_A}}{\|g_t\|_{A^{-1}}^2}
\end{aligned}$$

Hindsight minimization

- Focus on diagonal case (full matrix case similar)

$$\min_s \sum_{t=1}^{T} \left\langle g_t, \mathrm{diag}(s)^{-1} g_t \right\rangle \quad \text{subject to} \quad s \succeq 0, \langle 1, s \rangle \leq C$$

- Let $g_{1:T,j}$ be vector of jth component. Solution is of form

$$s_j \propto \left\| g_{1:T,j} \right\|_2$$

Low regret to the best A

$$\sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - f_t(x^*) - \varphi(x^*)$$

$$\leq \frac{1}{2\eta} \underbrace{\sum_{t=1}^{T} \left(\|x_t - x^*\|_{A_t}^2 - \|x_{t+1} - x^*\|_{A_t}^2 \right)}_{\text{Term I}} + \frac{\eta}{2} \underbrace{\sum_{t=1}^{T} \|g_t\|_{A_t^{-1}}^2}_{\text{Term II}}$$

Bounding Terms

Define $D_\infty = \max_t \|x_t - x^*\|_\infty \le \sup_{x \in \mathcal{X}} \|x - x^*\|_\infty$

- Term I:

$$\sum_{t=1}^{T} \left(\|x_t - x^*\|_{A_t}^2 - \|x_{t+1} - x^*\|_{A_t}^2 \right) \le D_\infty^2 \sum_{j=1}^{d} \|g_{1:T,j}\|_2$$

Bounding Terms

Define $D_\infty = \max_t \|x_t - x^*\|_\infty \le \sup_{x \in \mathcal{X}} \|x - x^*\|_\infty$

- Term I:

$$\sum_{t=1}^{T} \left(\|x_t - x^*\|_{A_t}^2 - \|x_{t+1} - x^*\|_{A_t}^2 \right) \le D_\infty^2 \sum_{j=1}^{d} \|g_{1:T,j}\|_2$$

- Term II:

$$\sum_{t=1}^{T} \|g_t\|_{A_t^{-1}}^2 \le 2 \sum_{t=1}^{T} \|g_t\|_{A_T^{-1}}^2 = 2 \sum_{j=1}^{d} \|g_{1:T,j}\|_2$$

$$= 2 \sqrt{\inf_s \left\{ \sum_{t=1}^{T} \langle g_t, \mathrm{diag}(s)^{-1} g_t \rangle \;\middle|\; s \succeq 0, \;\; \langle 1, s \rangle \le \sum_{j=1}^{d} \|g_{1:T,j}\|_2 \right\}}$$