



# EECS 559

## Optimization Methods for SIPML

### Lecture 5 – Stochastic Gradient Descent

Instructor: Prof. Qing Qu ([qingqu@umich.edu](mailto:qingqu@umich.edu))

2



## Lecture Agenda

- Method Introduction
- Convergence Analysis
- Stochastic Variance Reduced Gradient

## Stochastic Optimization

$$\min_{\mathbf{x}} F(\mathbf{x}) = \underbrace{\mathbb{E}_{\xi \sim \mathcal{D}} [f(\mathbf{x}; \xi)]}_{\text{expected/population risk}}$$

- $\xi$ : randomness in the problem,  $\mathcal{D}$ : the distribution of  $\xi$ .
- suppose  $f(\cdot, \xi)$  is *convex* for every  $\xi$ , hence  $F(\mathbf{x})$  is *convex*
- evaluating high-dimensional expectation is expensive.

3



## A Natural Solution

$$\min_{\mathbf{x}} F(\mathbf{x}) = \underbrace{\mathbb{E}_{\xi \sim \mathcal{D}} [f(\mathbf{x}; \xi)]}_{\text{expected/population risk}}$$

Under “mild” technical conditions:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k - \tau_k \nabla F(\mathbf{x}_k) = \mathbf{x}_k - \tau_k \nabla \mathbb{E}[f(\mathbf{x}_k; \xi)] \\ &= \mathbf{x}_k - \tau_k \mathbb{E}[\nabla_{\mathbf{x}} f(\mathbf{x}_k; \xi)] \end{aligned}$$

### Problems:

- The distribution of  $\xi$  might be *unknown*
- even if it is known, evaluating high-dimensional expectation is often difficult

4



## Solution: Empirical Risk Minimization (ERM)

Let  $\{\mathbf{a}_i, y_i\}_{i=1}^m$  be  $m$  random *i.i.d.* samples, and consider

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} F_m(\mathbf{x}) &= \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}), \quad f_i(\mathbf{x}) = h_i(\mathbf{x}) + \psi(\mathbf{x}) \\ &:= \frac{1}{m} \sum_{i=1}^m \underbrace{h(\mathbf{x}; \{\mathbf{a}_i, y_i\})}_{h_i(\mathbf{x}): \text{loss for } i\text{th sample}} + \underbrace{\psi(\mathbf{x})}_{\text{regularizer}}. \end{aligned}$$

5



## Empirical Risk Minimization (ERM)

$$\min_{\mathbf{x} \in \mathbb{R}^n} F_m(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x}), \quad f_i(\mathbf{x}) = h_i(\mathbf{x}) + \psi(\mathbf{x}).$$

- Linear regression:**  $h_i(\mathbf{x}) = \frac{1}{2} (\mathbf{a}_i^\top \mathbf{x} - y_i)^2$ ,  $\psi(\mathbf{x}) = 0$ ;
- Logistic regression:**  $h_i(\mathbf{x}) = \log(1 + e^{-y_i \mathbf{a}_i^\top \mathbf{x}})$ ,  $\psi(\mathbf{x}) = 0$ ;
- Lasso:**  $h_i(\mathbf{x}) = \frac{1}{2} (\mathbf{a}_i^\top \mathbf{x} - y_i)^2$ ,  $\psi(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$ ;
- SVM:**  $h_i(\mathbf{x}) = \max\{0, 1 - y_i \mathbf{a}_i^\top \mathbf{x}\}$ ,  $\psi(\mathbf{x}) = \frac{\lambda}{2} \|\mathbf{x}\|_2$ ;
- Training deep neural network...**

6



## Gradient Descent (GD)

- (Full batch) gradient descent:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \tau_k \cdot \nabla F_m(\mathbf{x}_k), \quad \nabla F_m(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m \nabla f_i(\mathbf{x})$$

when  $m$  is large, computing the gradient could be expensive: *linear* in the number of samples  $O(m)$ .

Overall, complexity of evaluating the gradient is  $O(mn)$ .

7



## Stochastic Gradient Descent (SGD)

$$\min_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{x}) := \mathbb{E}_{\xi \in \mathcal{D}} [f(\mathbf{x}; \xi)]$$

- The SGD in general can be written as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \tau_k \cdot \mathbf{g}(\mathbf{x}_k; \xi_k)$$

with  $\mathbf{g}(\mathbf{x}_k; \xi_k)$  being an *unbiased* estimation of  $\nabla F(\mathbf{x}_k)$

$$\mathbb{E}[\mathbf{g}(\mathbf{x}_k; \xi_k)] = \nabla F(\mathbf{x}_k)$$

8



## Stochastic Gradient Descent (SGD)

- Example of SGD:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \tau_k \cdot \mathbf{g}(\mathbf{x}_k), \quad \mathbf{g}(\mathbf{x}_k) = \frac{1}{b} \sum_{i \in B_k} \nabla f_i(\mathbf{x})$$

approximate  $\nabla F(\mathbf{x})$  by a random batch of samples  $B_k \subset [m]$  of a fixed size  $|B_k| \equiv b \ll m$ .

- It reduces the complexity to  $O(n)$ ;
- $\mathbf{g}(\mathbf{x})$  gives an unbiased estimator of  $\nabla F(\mathbf{x})$ :

$$\mathbb{E}[\mathbf{g}(\mathbf{x})] = \nabla F(\mathbf{x})$$

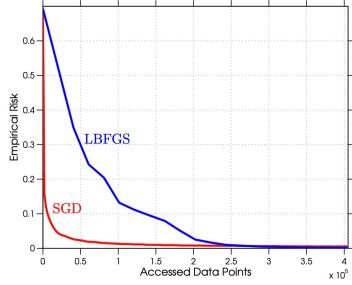
9

UNIVERSITY OF MICHIGAN

10

UNIVERSITY OF MICHIGAN

## SGD for Empirical Risk Minimization



Binary Classification with logistic loss and RCV1 data set

11

UNIVERSITY OF MICHIGAN

12

UNIVERSITY OF MICHIGAN

## Lecture Agenda

- Method Introduction
- **Convergence Analysis**
- Stochastic Variance Reduced Gradient

## Strongly Convex and Smooth Problems

$$\min_{\mathbf{x} \in \mathbb{R}^n} F := \mathbb{E}_{\xi \in \mathcal{D}} [f(\mathbf{x}; \xi)]$$

- $f$ :  $\mu$ -strongly convex,  $L$ -smooth;
- Given  $\{\xi_0, \dots, \xi_k\}$ ,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \tau_k \cdot \mathbf{g}(\mathbf{x}_k; \xi_k)$$

with  $\mathbf{g}(\mathbf{x}_k; \xi_k)$  being an unbiased estimate of  $\nabla F(\mathbf{x}_k)$ .

- For all  $\mathbf{x}$ ,

$$\mathbb{E} [\|\mathbf{g}(\mathbf{x}; \xi)\|_2^2] \leq \sigma_g^2 + c_g \|\nabla F(\mathbf{x})\|_2^2$$

13

UNIVERSITY OF MICHIGAN

## Convergence of SGD with Fixed Stepsizes

**Theorem.** Assume that  $f$  is  $\mu$ -strongly convex and  $L$ -smooth, and

$$\mathbb{E} [\|\mathbf{g}(\mathbf{x}; \xi)\|_2^2] \leq \sigma_g^2 + c_g \|\nabla F(\mathbf{x})\|_2^2.$$

If the stepsize  $\tau_k \equiv \tau \leq 1/(Lc_g)$ , then SGD achieves

$$F(\mathbf{x}_k) - F(\mathbf{x}_*) \leq \underbrace{(1 - \tau\mu)^k}_{\text{linear convergence}} (F(\mathbf{x}_0) - F(\mathbf{x}_*)) + \frac{\tau L \sigma_g^2}{2\mu}$$

Check Theorem 4.6 of Bottou, Curtis, Nocedal'18 for the proof.

Theorem 4.6 in Léon Bottou, Frank E. Curtis, Jorge Nocedal [Optimization Methods for Large-Scale Machine Learning](#), 2018

14

UNIVERSITY OF MICHIGAN

## Implications: SGD with Fixed Stepsizes

$$F(\mathbf{x}_k) - F(\mathbf{x}_*) \leq \underbrace{(1 - \tau\mu)^k}_{\text{linear convergence}} (F(\mathbf{x}_0) - F(\mathbf{x}_*)) + \frac{\tau L \sigma_g^2}{2\mu}$$

- Fast (linear!) convergence at the very beginning
- Converges to some neighborhood of  $\mathbf{x}_*$  due to

$$\mathbb{E} [\|\mathbf{g}(\mathbf{x}; \xi)\|_2^2] \leq \sigma_g^2 + c_g \|\nabla F(\mathbf{x})\|_2^2$$

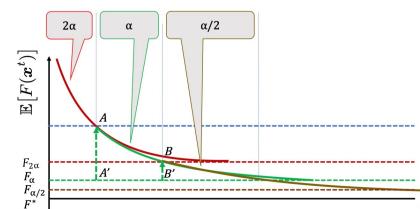
- when  $\sigma_g = 0$ , it converges linearly to optimal points
- smaller stepsizes  $\tau$  yield better converging points

15

UNIVERSITY OF MICHIGAN

## One Practical Strategy

Run SGD with fixed stepsizes; whenever progress stalls, reduce stepsizes and continue SGD



whenever progress stalls, we half the stepsizes and repeat

16

UNIVERSITY OF MICHIGAN

## Convergence with Diminishing Stepsizes

**Theorem. (Convergence of SGD with Diminishing Stepsizes)** Suppose  $F$  is  $\mu$ -strongly convex, and the following

$$\mathbb{E} [\|\mathbf{g}(\mathbf{x}; \boldsymbol{\xi})\|_2^2] \leq \sigma_g^2 + c_g \|\nabla F(\mathbf{x})\|_2^2$$

holds with  $c_g = 0$ . If  $\tau_k = \frac{\theta}{k+1}$  for some  $\theta > \frac{1}{2\mu}$ , then

$$\mathbb{E} [\|\mathbf{x}_k - \mathbf{x}_*\|_2^2] \leq \frac{c_\theta}{k+1},$$

where  $c_\theta = \max \left\{ \frac{2\theta^2\sigma_g^2}{2\mu\theta-1}, \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2 \right\}$ .

Theorem 4.7 in Léon Bottou, Frank E. Curtis, Jorge Nocedal [Optimization Methods for Large-Scale Machine Learning](#), 2018

17



## Optimality of Convergence Rate

**Theorem. (Informal, Nemirovski & Yudin'83)**

For strongly convex functions, no algorithm performing  $k$  queries to noisy first-order oracles can achieve an accuracy better than  $O(1/k)$ .

- SGD with stepsize  $\tau_k$  can achieve optimal order of convergence rate  $O(1/k)$ .

"Problem complexity and method efficiency in optimization," A. Nemirovski, D. Yudin, Wiley, 1983.

18



## Comparing SGD with GD

	Iter. Complexity	Per-Iter Cost	Total Comput. Cost
GD	$O(\log \frac{1}{\varepsilon})$	$O(mn)$	$O(mn \log \frac{1}{\varepsilon})$
SGD	$O(\frac{1}{\varepsilon})$	$O(n)$	$O(\frac{n}{\varepsilon})$

SGD is more appealing than GD for *large m* and *moderate accuracy*  $\varepsilon$  (whenever  $\frac{1}{\varepsilon} < m \log \frac{1}{\varepsilon}$ ).

➤ it often arises in the *big data* regime in machine learning !

19



## General Convex Problems

What if we lose *strong* convexity?

$$\min_{\mathbf{x} \in \mathbb{R}^n} F := \mathbb{E}_{\boldsymbol{\xi} \in \mathcal{D}} [f(\mathbf{x}; \boldsymbol{\xi})]$$

- $f$ : convex,  $L$ -smooth;
- Given  $\{\boldsymbol{\xi}_0, \dots, \boldsymbol{\xi}_k\}$ ,  $\mathbf{x}_{k+1} = \mathbf{x}_k - \tau_k \cdot \mathbf{g}(\mathbf{x}_k; \boldsymbol{\xi}_k)$  with  $\mathbf{g}(\mathbf{x}_k; \boldsymbol{\xi}_k)$  being an *unbiased* estimate of  $\nabla F(\mathbf{x}_k)$ .
- For all  $\mathbf{x}$ ,  $\mathbb{E} [\|\mathbf{g}(\mathbf{x}; \boldsymbol{\xi})\|_2^2] \leq \sigma_g^2 + c_g \|\nabla F(\mathbf{x})\|_2^2$

20



## General Convex Problems

Suppose we return a weighted average

$$\tilde{\mathbf{x}}_k := \sum_{i=0}^k \frac{\tau_i}{\sum_{j=0}^k \tau_j} \mathbf{x}_i, \quad \mathbf{x}_{i+1} = \mathbf{x}_i - \tau_i \cdot \mathbf{g}(\mathbf{x}_i; \boldsymbol{\xi}_i)$$

**Theorem. (Convergence for Convex Problems)**

Under the assumptions on the previous page, one has

$$\mathbb{E} [F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}_*)] \leq \frac{1}{2 \sum_{i=0}^k \tau_i} \left( \mathbb{E} [\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2] + \sigma_g^2 \sum_{i=0}^k \tau_i^2 \right).$$

Zeyuan Allen-Zhu, Elad Hazan, Optimal Black-Box Reductions Between Optimization Objectives, NeurIPS'16

21



## General Convex Problems

**Theorem. (Convergence for Convex Problems)**

Under the assumptions on the previous page, one has

$$\mathbb{E} [F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}_*)] \leq \frac{1}{2 \sum_{i=0}^k \tau_i} \left( \mathbb{E} [\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2] + \sigma_g^2 \sum_{i=0}^k \tau_i^2 \right).$$

If  $\tau_k \asymp 1/\sqrt{k}$ , then

$$\mathbb{E} [F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}_*)] \leq O\left(\frac{\log k}{\sqrt{k}}\right)$$

Zeyuan Allen-Zhu, Elad Hazan, Optimal Black-Box Reductions Between Optimization Objectives, NeurIPS'16

22



## Lecture Agenda

- Method Introduction
- Convergence Analysis
- Stochastic Variance Reduced Gradient**

## Vanilla SGD

For  $\mu$ -strongly convex problem, what we have learned so far:

- SGD with (large) constant stepsize  $\tau_k \in O(1)$  tends to oscillate around global minimum:

$$F(\mathbf{x}_k) - F(\mathbf{x}_*) \leq (1 - \tau\mu)^k (F(\mathbf{x}_0) - F(\mathbf{x}_*)) + \frac{\tau L \sigma_g^2}{2\mu}$$

- choosing conservative  $\tau_k \in O(1/k)$  mitigates oscillation, but results in slow convergence

$$\mathbb{E} [\|\mathbf{x}_k - \mathbf{x}_*\|_2^2] \leq \frac{c_\theta}{k+1}.$$

23



24



## Design $\mathbf{g}_k$ with Reduced Variability $\sigma_g^2$ ?

$$F(\mathbf{x}_k) - F(\mathbf{x}_*) \leq (1 - \tau\mu)^k (F(\mathbf{x}_0) - F(\mathbf{x}_*)) + \frac{\tau L \sigma_g^2}{2\mu}$$

- vanilla SGD:  $\mathbf{g}_k = \nabla f_{i_k}(\mathbf{x}_k)$ ,  $i_k \sim \mathcal{U}([m])$   
 $\sigma_g^2$  is non-negligible even when  $\mathbf{x}_k = \mathbf{x}_*$ ;
- **question:** design  $\mathbf{g}_k$  with reduced variability  $\sigma_g^2$  ?

25

UNIVERSITY OF MICHIGAN

## Design $\mathbf{g}_k$ with Reduced Variability $\sigma_g^2$ ?

$$F(\mathbf{x}_k) - F(\mathbf{x}_*) \leq (1 - \tau\mu)^k (F(\mathbf{x}_0) - F(\mathbf{x}_*)) + \frac{\tau L \sigma_g^2}{2\mu}$$

- **question:** design  $\mathbf{g}_k$  with reduced variability  $\sigma_g^2$  ?
  - **idea:** take some  $\mathbf{v}_k$  that (i)  $\langle \mathbf{v}_k, \nabla f_{i_k}(\mathbf{x}_k) \rangle > 0$ , (ii)  $\mathbb{E}[\mathbf{v}_k] = \mathbf{0}$
- $\mathbf{g}_k = \nabla f_{i_k}(\mathbf{x}_k) - \mathbf{v}_k$
- use historical gradient information to produce such a  $\mathbf{v}_k$   
 ➤ so  $\mathbf{g}_k$  is still an unbiased estimate of  $\nabla F(\mathbf{x}_k)$  with reduced  $\sigma_g^2$

26

UNIVERSITY OF MICHIGAN

## Stochastic Variance Reduced Gradient (SVRG)

$$\min_{\mathbf{x} \in \mathbb{R}^n} F_m(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{x})$$

- $f_i$ : convex and  $L$ -smooth, no regularizations
- $F_m$ :  $\mu$ -strongly convex
- $\kappa := L/\mu$ : condition number

27

UNIVERSITY OF MICHIGAN

## Stochastic Variance Reduced Gradient (SVRG)

**Key idea:** if we have access to a historical *anchor point*  $\mathbf{x}^a$  and  $\nabla F_m(\mathbf{x}^a)$ , then

$$\underbrace{\nabla f_{i_k}(\mathbf{x}_k) - \nabla f_{i_k}(\mathbf{x}^a)}_{\rightarrow \mathbf{0} \text{ if } \mathbf{x}_k \approx \mathbf{x}^a} + \underbrace{\nabla F_m(\mathbf{x}^a)}_{\rightarrow \mathbf{0} \text{ if } \mathbf{x}^a \approx \mathbf{x}_*}, \quad \text{with } i_k \sim \mathcal{U}([m]).$$

- is an *unbiased estimator* of  $\nabla F(\mathbf{x}_k)$
- converges to  $\mathbf{0}$  when  $\mathbf{x}_k \approx \mathbf{x}_a \approx \mathbf{x}_*$

28

UNIVERSITY OF MICHIGAN

## Stochastic Variance Reduced Gradient (SVRG)

- Operate in *epochs*
- In the  $k$ -th epoch:
  1. **very beginning:** record the anchor point  $\mathbf{x}_k^a$ , and compute the *full batch* gradient  $\nabla F_m(\mathbf{x}_k^a)$
  2. **inner loop:** use the anchor point  $\mathbf{x}_k^a$  to reduce the variance
$$\mathbf{x}_{k,t+1} = \mathbf{x}_{k,t} - \tau [\nabla f_{i_{k,t}}(\mathbf{x}_{k,t}) - \nabla f_{i_{k,t}}(\mathbf{x}_k^a) + \nabla F_m(\mathbf{x}_k^a)].$$

**A hybrid approach:** the *full batch* gradient is computed *only once* per epoch

29

UNIVERSITY OF MICHIGAN

## Algorithm Pipeline for SVRG

---

**Algorithm 1** SVRG for finite-sum optimization

---

```

for  $k = 1, 2, \dots$  do
     $\mathbf{x}_k^a \leftarrow \mathbf{x}_{k-1,N}$  and compute  $\underbrace{\nabla F_m(\mathbf{x}_k^a)}_{\text{full batch gradient}}$ 
    initialize  $\mathbf{x}_{k,0} \leftarrow \mathbf{x}_k^a$ 
    for  $t = 0, \dots, N-1$  do
         $N$  inner loops for each epoch
        draw  $i_{k,t}$  uniformly from  $[m]$  and compute
         $\mathbf{x}_{k,t+1} = \mathbf{x}_{k,t} - \tau [\underbrace{\nabla f_{i_{k,t}}(\mathbf{x}_{k,t}) - \nabla f_{i_{k,t}}(\mathbf{x}_k^a)}_{\text{stochastic gradient}} + \nabla F_m(\mathbf{x}_k^a)]$ 
    end for
end for

```

---

30

UNIVERSITY OF MICHIGAN

## Remarks for SVRG

- We use constant stepsize  $\tau$
- Each epoch contains  $2N+m$  gradient computations of  $f_i$ 
  - the full batch gradient is computed only once every  $N$  iterations;
  - the average per-iteration cost of SVRG is comparable to that of SGD if  $N \asymp m$ .

31

UNIVERSITY OF MICHIGAN

## Convergence Analysis of SVRG

**Theorem (Johnson & Zhang'13)**

Assume each  $f_i$  is convex and  $L$ -smooth and  $F$  is  $\mu$ -strongly convex. Choose the inner loop  $N$  large enough such that

$$\rho = \frac{1}{\mu\tau(1-2\tau L)N} + \frac{2\tau L}{1-2\tau L} < 1,$$

then we have

$$\mathbb{E}[F_m(\mathbf{x}_k^a) - F_m(\mathbf{x}_*)] \leq \rho^k \cdot (F_m(\mathbf{x}_0^a) - F(\mathbf{x}_*)).$$

**Linear convergence:** choose  $N \geq L/\mu = \kappa$  and stepsize  $\tau \in O(1/L)$  so that  $0 < \rho < 1/2$  and  $O(\log \frac{1}{\epsilon})$  epochs to attain  $\epsilon$ -accuracy.

See also **Theorem 5.1** in Léon Bottou, Frank E. Curtis, Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning, 2018

32

UNIVERSITY OF MICHIGAN

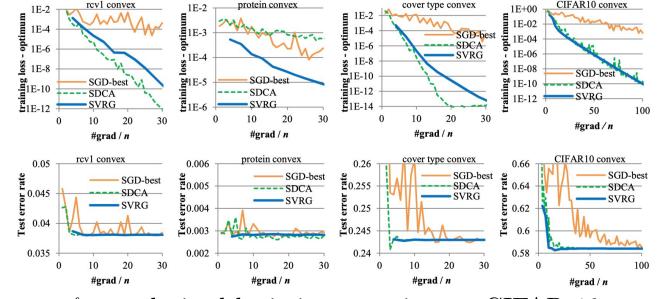
## Comparisons with GD and SGD

	SVRG	GD	SGD
Comp. Cost	$(n + \kappa) \log \frac{1}{\varepsilon}$	$n\kappa \log \frac{1}{\varepsilon}$	$\frac{\kappa^2}{\varepsilon}$

33

UNIVERSITY OF MICHIGAN

## Numerical Example: Logistic Regression



$\ell_2$ -regularized logistic regression on CIFAR-10

34

UNIVERSITY OF MICHIGAN

## Further Readings

- *Optimization Methods for Large-scale Machine Learning*. Léon Bottou, Frank E. Curtis, and Jorge Nocedal. SIAM Review, 2018.
- *Convex Optimization: Algorithms & Complexity*. Sébastien Bubeck, Foundation & Trends in Machine Learning (Chapter 6).
- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. John Wright, Yi Ma. (Chapter 8.6)

35

UNIVERSITY OF MICHIGAN

## Further Readings

- *Problem Complexity and Method Efficiency in Optimization*. Arkadi Nemirovski and David Yudin, Wiley, 1983.
- *Accelerating Stochastic Gradient Descent using Predictive Variance Reduction*. Rie Johnson and Tong Zhang, NeurIPS'13, 2013.
- *A Proximal Stochastic Gradient Method with Progressive Variance Reduction*. Lin Xiao and Tong Zhang, SIAM Journal on Optimization, 2014.
- *Recent Advances in Stochastic Convex & Nonconvex Optimization*. Zeyuan Allen-Zhu, ICML Tutorial, 2017.  
<http://people.csail.mit.edu/zeyuan/topics/icml-2017>

36

UNIVERSITY OF MICHIGAN