

# Optimization for Machine Learning

## CS-439

Lecture 4: Proximal and Subgradient Descent

**Nicolas Flammarion**

EPFL – [github.com/epfml/OptML\\_course](https://github.com/epfml/OptML_course)

March 13, 2020

## Section 3.6

# Proximal Gradient Descent

# Composite optimization problems

Consider objective functions composed as

$$f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x})$$

where  $g$  is a “nice” function, where as  $h$  is a “simple” additional term, which however doesn't satisfy the assumptions of niceness which we used in the convergence analysis so far.

In particular, an important case is when  $h$  is not differentiable.

## Idea

The classical gradient step for minimizing  $g$ :

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{y}} g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top (\mathbf{y} - \mathbf{x}_t) + \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{x}_t\|^2 .$$

For the stepsize  $\gamma := \frac{1}{L}$  it exactly minimizes the local quadratic model of  $g$  at our current iterate  $\mathbf{x}_t$ , formed by the smoothness property with parameter  $L$ .

Now for  $f = g + h$ , keep the same for  $g$ , and add  $h$  unmodified.

$$\begin{aligned} \mathbf{x}_{t+1} &:= \operatorname{argmin}_{\mathbf{y}} g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top (\mathbf{y} - \mathbf{x}_t) + \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{x}_t\|^2 + h(\mathbf{y}) \\ &= \operatorname{argmin}_{\mathbf{y}} \frac{1}{2\gamma} \|\mathbf{y} - (\mathbf{x}_t - \gamma \nabla g(\mathbf{x}_t))\|^2 + h(\mathbf{y}) , \end{aligned}$$

the **proximal gradient descent** update.

# The proximal gradient descent algorithm

An iteration of **proximal gradient descent** is defined as

$$\mathbf{x}_{t+1} := \text{prox}_{h,\gamma}(\mathbf{x}_t - \gamma \nabla g(\mathbf{x}_t)) .$$

where the **proximal mapping** for a given function  $h$ , and parameter  $\gamma > 0$  is defined as

$$\text{prox}_{h,\gamma}(\mathbf{z}) := \underset{\mathbf{y}}{\text{argmin}} \left\{ \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{z}\|^2 + h(\mathbf{y}) \right\} .$$

The update step can be equivalently written as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma G_\gamma(\mathbf{x}_t)$$

for  $G_{h,\gamma}(\mathbf{x}) := \frac{1}{\gamma} \left( \mathbf{x} - \text{prox}_{h,\gamma}(\mathbf{x} - \gamma \nabla g(\mathbf{x})) \right)$  being the so called **generalized gradient** of  $f$ .

# A generalization of gradient descent?

- ▶  $h \equiv 0$ : recover **gradient descent**
- ▶  $h \equiv \iota_X$ : recover **projected gradient descent!**

Given a closed convex set  $X$ , the **indicator function** of the set  $X$  is given as the convex function

$$\begin{aligned} \iota_X : \mathbb{R}^d &\rightarrow \mathbb{R} \cup +\infty \\ \mathbf{x} \mapsto \iota_X(\mathbf{x}) &:= \begin{cases} 0 & \text{if } \mathbf{x} \in X, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

Proximal mapping becomes

$$\text{prox}_{h,\gamma}(\mathbf{z}) := \underset{\mathbf{y}}{\text{argmin}} \left\{ \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{z}\|^2 + \iota_X(\mathbf{y}) \right\} = \underset{\mathbf{y} \in X}{\text{argmin}} \|\mathbf{y} - \mathbf{z}\|^2$$

## Convergence in $\mathcal{O}(1/\varepsilon)$ steps

Same as vanilla case for smooth functions, but now for any  $h$  for which we can compute the proximal mapping.

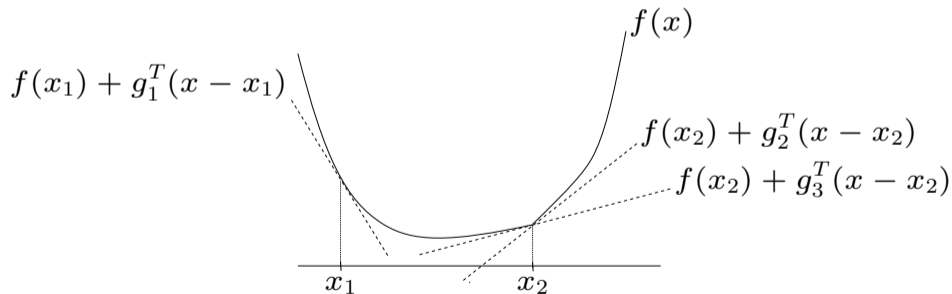
# Subgradients

What if  $f$  is not differentiable?

## Definition

$\mathbf{g} \in \mathbb{R}^d$  is a **subgradient** of  $f$  at  $\mathbf{x}$  if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top (\mathbf{y} - \mathbf{x}) \quad \text{for all } \mathbf{y} \in \text{dom}(f)$$

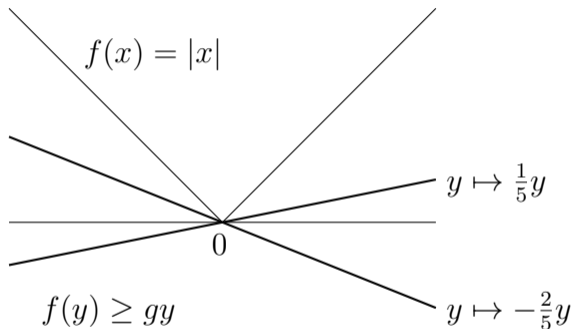


$\partial f(\mathbf{x}) \subseteq \mathbb{R}^d$  is the **subdifferential**, the set of subgradients of  $f$  at  $\mathbf{x}$ .



## Subgradients II

Example:



Subgradient condition at  $x = 0$ :  $f(y) \geq f(0) + g(y - 0) = gy$ .

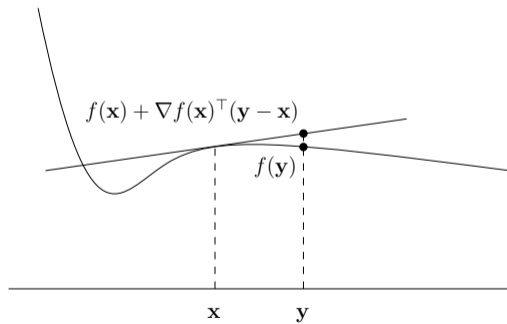
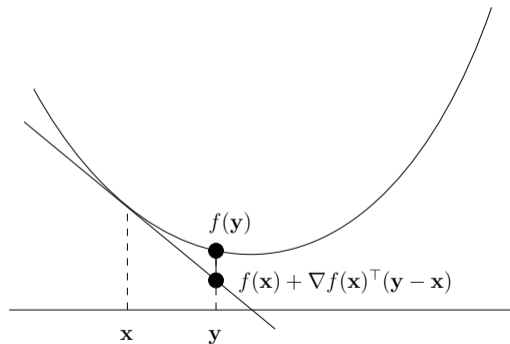
$$\partial f(0) = [-1, 1]$$

# Subgradients III

Lemma (Exercise 23)

If  $f : \text{dom}(f) \rightarrow \mathbb{R}$  is differentiable at  $\mathbf{x} \in \text{dom}(f)$ , then  $\partial f(\mathbf{x}) \subseteq \{\nabla f(\mathbf{x})\}$ .

Either exactly one subgradient  $\nabla f(\mathbf{x})$ ... ..or no subgradient at all.

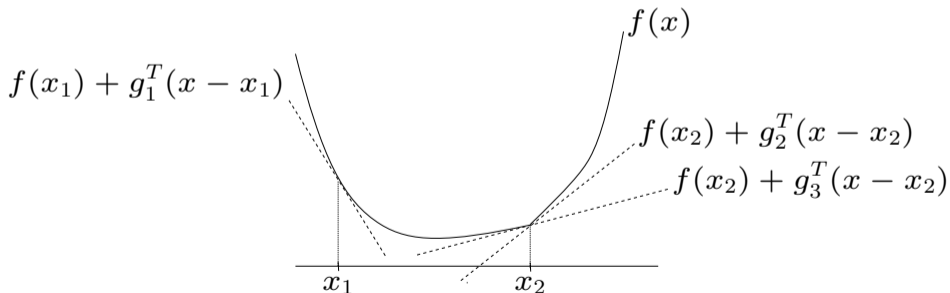


# Subgradient characterization of convexity

“convex = subgradients everywhere”

Lemma (Exercise 24)

A function  $f : \mathbf{dom}(f) \rightarrow \mathbb{R}$  is convex if and only if  $\mathbf{dom}(f)$  is convex and  $\partial f(\mathbf{x}) \neq \emptyset$  for all  $\mathbf{x} \in \mathbf{dom}(f)$ .



# Convex and Lipschitz = bounded subgradients

## Lemma (Exercise 25)

Let  $f : \mathbf{dom}(f) \rightarrow \mathbb{R}$  be convex,  $\mathbf{dom}(f)$  open,  $B \in \mathbb{R}_+$ . Then the following two statements are equivalent.

- (i)  $\|\mathbf{g}\| \leq B$  for all  $\mathbf{x} \in \mathbf{dom}(f)$  and all  $\mathbf{g} \in \partial f(\mathbf{x})$ .
- (ii)  $|f(\mathbf{x}) - f(\mathbf{y})| \leq B\|\mathbf{x} - \mathbf{y}\|$  for all  $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$ .

# Subgradient optimality condition

## Lemma

Suppose that  $f : \text{dom}(f) \rightarrow \mathbb{R}$  and  $\mathbf{x} \in \text{dom}(f)$ . If  $\mathbf{0} \in \partial f(\mathbf{x})$ , then  $\mathbf{x}$  is a global minimum.

## Proof.

By definition of subgradients,  $\mathbf{g} = \mathbf{0} \in \partial f(\mathbf{x})$  gives

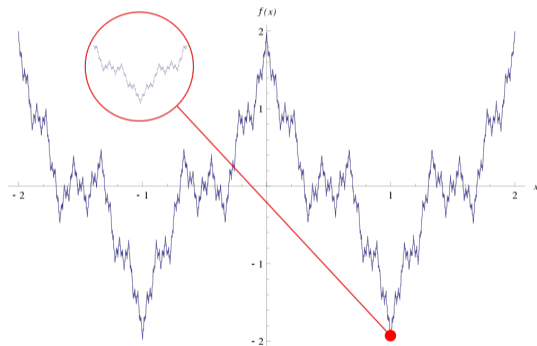
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top (\mathbf{y} - \mathbf{x}) = f(\mathbf{x})$$

for all  $\mathbf{y} \in \text{dom}(f)$ , so  $\mathbf{x}$  is a global minimum. □

# Differentiability of convex functions

How “wild” can a non-differentiable convex function be?

Weierstrass function: a function that is continuous **everywhere** but differentiable **nowhere**



<https://commons.wikimedia.org/wiki/File:WeierstrassFunction.svg>

# Differentiability of convex functions

Theorem ([Roc97, Theorem 25.5])

A *convex* function  $f : \mathbf{dom}(f) \rightarrow \mathbb{R}$  is differentiable *almost everywhere*.

In other words:

- ▶ Set of points where  $f$  is non-differentiable has measure 0 (no volume).
- ▶ For all  $\mathbf{x} \in \mathbf{dom}(f)$  and all  $\varepsilon > 0$ , there is a point  $\mathbf{x}'$  such that  $\|\mathbf{x} - \mathbf{x}'\| < \varepsilon$  and  $f$  is differentiable at  $\mathbf{x}'$ .

# The subgradient descent algorithm

**Subgradient descent:** choose  $\mathbf{x}_0 \in \mathbb{R}^d$ .

$$\text{Let } \mathbf{g}_t \in \partial f(\mathbf{x}_t)$$

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}_t$$

for **times**  $t = 0, 1, \dots$ , and **stepsizes**  $\gamma_t \geq 0$ .

Stepsize can vary with time!

This is possible in (projected) gradient descent as well, but so far, we didn't need it.



## Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

### Theorem

Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be convex and  $B$ -Lipschitz continuous with a global minimum  $\mathbf{x}^*$ ; furthermore, suppose that  $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq R$ . Choosing the constant stepsize

$$\gamma := \frac{R}{B\sqrt{T}},$$

subgradient descent yields

$$\frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{RB}{\sqrt{T}}.$$

Proof is identical to the one of Theorem 2.1, except...

- ▶ In vanilla analysis, now use  $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$  instead of  $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$ .
- ▶ Inequality  $f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*)$  now follows from subgradient property instead of first-order characterization of convexity.

# Optimality of first-order methods

With all the convergence rates we have seen so far, a very natural question to ask is if these rates are best possible or not. Surprisingly, the rate can indeed not be improved in general.

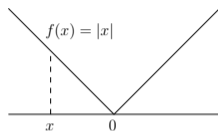
## Theorem (Nesterov)

*For any  $T \leq d - 1$  and starting point  $\mathbf{x}_0$ , there is a function  $f$  in the problem class of  $B$ -Lipschitz functions over  $\mathbb{R}^d$ , such that any (sub)gradient method has an objective error at least*

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \geq \frac{RB}{2(1 + \sqrt{T + 1})} .$$

## Smooth (non-differentiable) functions?

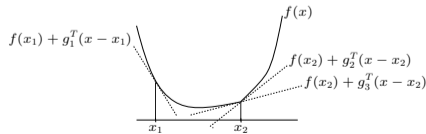
They don't exist (Exercise 26)!



At 0, graph can't be below a tangent paraboloid.

Can we still improve over  $O(1/\varepsilon^2)$  steps for Lipschitz functions?

Yes, if we also require strong convexity (graph is above not too flat tangent paraboloids).



# Strongly convex functions

## “Not too flat”

Straightforward generalization to the non-differentiable case:

### Definition

Let  $f : \mathbf{dom}(f) \rightarrow \mathbb{R}$  be convex,  $\mu \in \mathbb{R}_+, \mu > 0$ . Function  $f$  is called **strongly convex** (with parameter  $\mu$ ) if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f), \forall \mathbf{g} \in \partial f(\mathbf{x}).$$

# Strongly convex functions: characterization via “normal” convexity

## Lemma (Exercise 28)

Let  $f : \mathbf{dom}(f) \rightarrow \mathbb{R}$  be convex,  $\mathbf{dom}(f)$  open,  $\mu \in \mathbb{R}_+, \mu > 0$ .  $f$  is strongly convex with parameter  $\mu$  if and only if  $f_\mu : \mathbf{dom}(f) \rightarrow \mathbb{R}$  defined by

$$f_\mu(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2} \|\mathbf{x}\|^2, \quad \mathbf{x} \in \mathbf{dom}(f)$$

is convex.

## Tame strong convexity

For fast convergence, we consider **additional** assumptions.

Smoothness? - Not an option in the non-differentiable case (Exercise 26).

Instead: assume that all subgradients  $\mathbf{g}_t$  that we encounter during the algorithm are bounded in norm.

May be realistic if...

- ▶ we start close to optimality
- ▶ we run **projected** subgradient descent over a compact set  $X$

May also fail!

- ▶ Over  $\mathbb{R}^d$ , strong convexity and bounded subgradients contradict each other! (Exercise 30).

## Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps

### Theorem

Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be strongly convex with parameter  $\mu > 0$  and let  $\mathbf{x}^*$  be the unique global minimum of  $f$ . With decreasing step size

$$\gamma_t := \frac{2}{\mu(t+1)}, \quad t > 0,$$

subgradient descent yields

$$f\left(\frac{2}{T(T+1)} \sum_{t=1}^T t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^*) \leq \frac{2B^2}{\mu(T+1)},$$

where  $B = \max_{t=1}^T \|\mathbf{g}_t\|$ .

↑

convex combination of iterates

## Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps II

Proof.

Vanilla analysis ( $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ ):

$$\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*) = \frac{\gamma_t}{2} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma_t} (\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2).$$

Lower bound from **strong** convexity:

$$\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*) \geq f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\mu}{2} \|\mathbf{x}_t - \mathbf{x}^*\|^2.$$

Putting it together (with  $\|\mathbf{g}_t\|^2 \leq B^2$ ):

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2} \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \frac{\gamma_t^{-1}}{2} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2.$$

Summing over  $t = 1, \dots, T$ : we used to have telescoping ( $\gamma_t = \gamma, \mu = 0$ )...



## Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps III

Proof.

So far we have:

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2} \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \frac{\gamma_t^{-1}}{2} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2.$$

To get telescoping, we would need  $\gamma_t^{-1} = \gamma_{t+1}^{-1} - \mu$ .

Works with  $\gamma_t^{-1} = \mu(1+t)$ , but **not**  $\gamma_t^{-1} = \mu(1+t)/2$  (the choice here).

Exercise 31: what happens with  $\gamma_t^{-1} = \mu(1+t)$ ?

Now: what happens with  $\gamma_t^{-1} = \mu(1+t)/2$  (the choice here)?

## Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps IV

Proof.

So far we have:

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2} \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \frac{\gamma_t^{-1}}{2} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2.$$

Plug in  $\gamma_t^{-1} = \mu(1+t)/2$  and multiply with  $t$  on both sides:

$$\begin{aligned} t \cdot (f(\mathbf{x}_t) - f(\mathbf{x}^*)) &\leq \frac{B^2t}{\mu(t+1)} + \frac{\mu}{4} \left( t(t-1) \|\mathbf{x}_t - \mathbf{x}^*\|^2 - (t+1)t \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \right) \\ &\leq \frac{B^2}{\mu} + \frac{\mu}{4} \left( t(t-1) \|\mathbf{x}_t - \mathbf{x}^*\|^2 - (t+1)t \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \right). \end{aligned}$$

## Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps **V**

Proof.

We have

$$\begin{aligned} t \cdot (f(\mathbf{x}_t) - f(\mathbf{x}^*)) &\leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4} \left( t(t-1) \|\mathbf{x}_t - \mathbf{x}^*\|^2 - (t+1)t \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \right) \\ &\leq \frac{B^2}{\mu} + \frac{\mu}{4} \left( t(t-1) \|\mathbf{x}_t - \mathbf{x}^*\|^2 - (t+1)t \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \right). \end{aligned}$$

Now we get telescoping...

$$\sum_{t=1}^T t \cdot (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \frac{TB^2}{\mu} + \frac{\mu}{4} \left( 0 - T(T+1) \|\mathbf{x}_{T+1} - \mathbf{x}^*\|^2 \right) \leq \frac{TB^2}{\mu}.$$

## Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps VI

Proof.

Almost done:

$$\underline{\sum_{t=1}^T t \cdot (f(\mathbf{x}_t) - f(\mathbf{x}^*))} \leq \frac{TB^2}{\mu} + \frac{\mu}{4} \left( 0 - T(T+1) \|\mathbf{x}_{T+1} - \mathbf{x}^*\|^2 \right) \leq \frac{TB^2}{\mu}.$$

Since

$$\frac{2}{T(T+1)} \sum_{t=1}^T t = 1,$$

Jensen's inequality yields

$$f\left(\frac{2}{T(T+1)} \sum_{t=1}^T t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^*) \leq \frac{2}{T(T+1)} \underline{\sum_{t=1}^T t \cdot (f(\mathbf{x}_t) - f(\mathbf{x}^*))}.$$

□

## Tame strong convexity: Discussion

$$f\left(\frac{2}{T(T+1)} \sum_{t=1}^T t \cdot \mathbf{x}_t\right) - f(\mathbf{x}^*) \leq \frac{2B^2}{\mu(T+1)},$$

Weighted average of iterates achieves the bound (later iterates have more weight)

Bound is independent of initial distance  $\|\mathbf{x}_0 - \mathbf{x}^*\| \dots$

$\dots$  but not really:  $B$  typically depends on  $\|\mathbf{x}_0 - \mathbf{x}^*\|$  (for example,  $B = \mathcal{O}(\|\mathbf{x}_0 - \mathbf{x}^*\|)$  for quadratic functions)

Recall: we can only hope that  $B$  is small (can be checked while running the algorithm)

What if we don't know the parameter  $\mu$  of strong convexity?

→ **Bad luck!** In practice, try some  $\mu$ 's, pick best solution obtained

# Bibliography



R. Tyrrell Rockafellar.

*Convex Analysis.*

Princeton Landmarks in Mathematics. Princeton University Press, 1997.