BOSTON UNIVERSITY

Lecture I: Adaptive Gradient Descent

ALINA ENE

ADFOCS ’21: Convex Optimization and Graph Algorithms

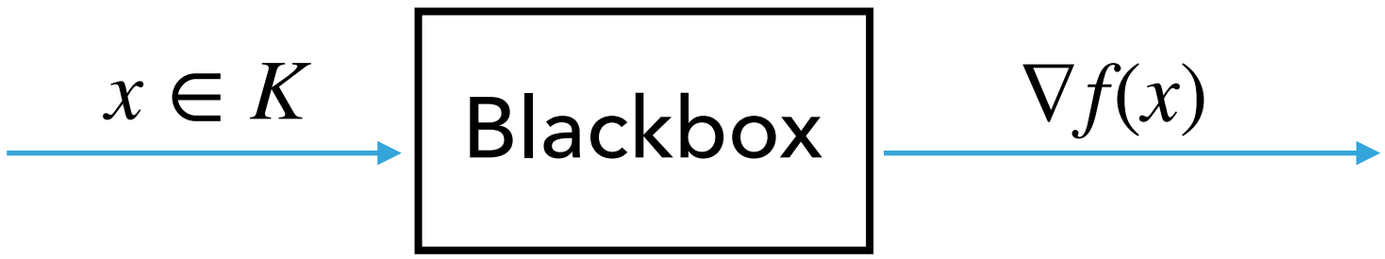
Problem Definition

$f\colon \mathbb{R}^d \to \mathbb{R}$ convex and differentiable function

$K \subseteq \mathbb{R}^d$ convex constraint set

$$\min_{x \in K} f(x)$$

Computational model: function access via first-order oracle



Goal: minimize number of queries $x_1, x_2, \ldots, x_T$ to obtain

$$f(x_{\text{out}}) - f(x^*) \leq \epsilon$$

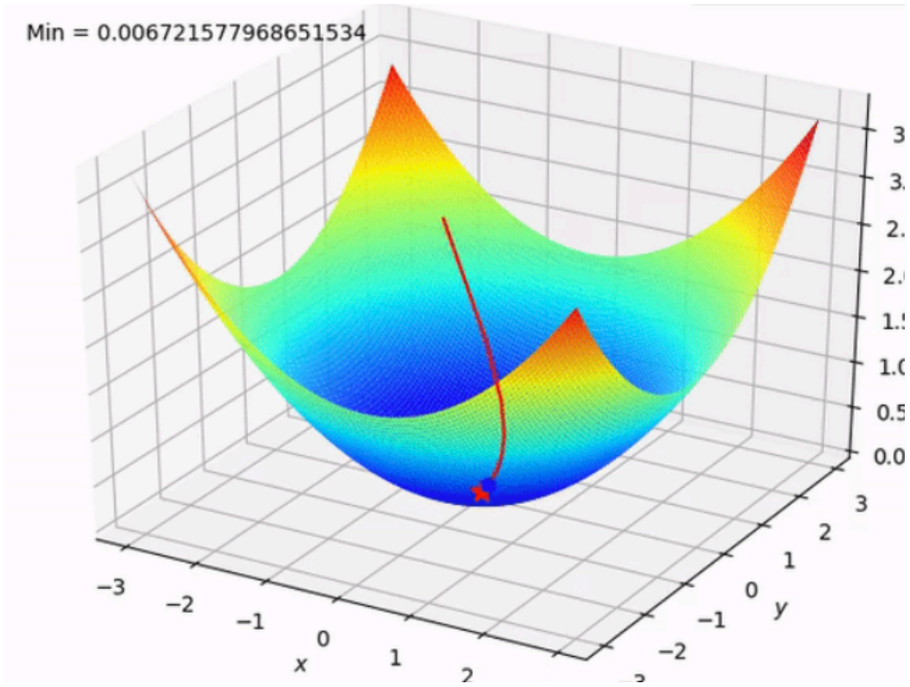
How to Optimize

$$\min_{x \in \mathbb{R}^d} f(x)$$

Gradient Descent

$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

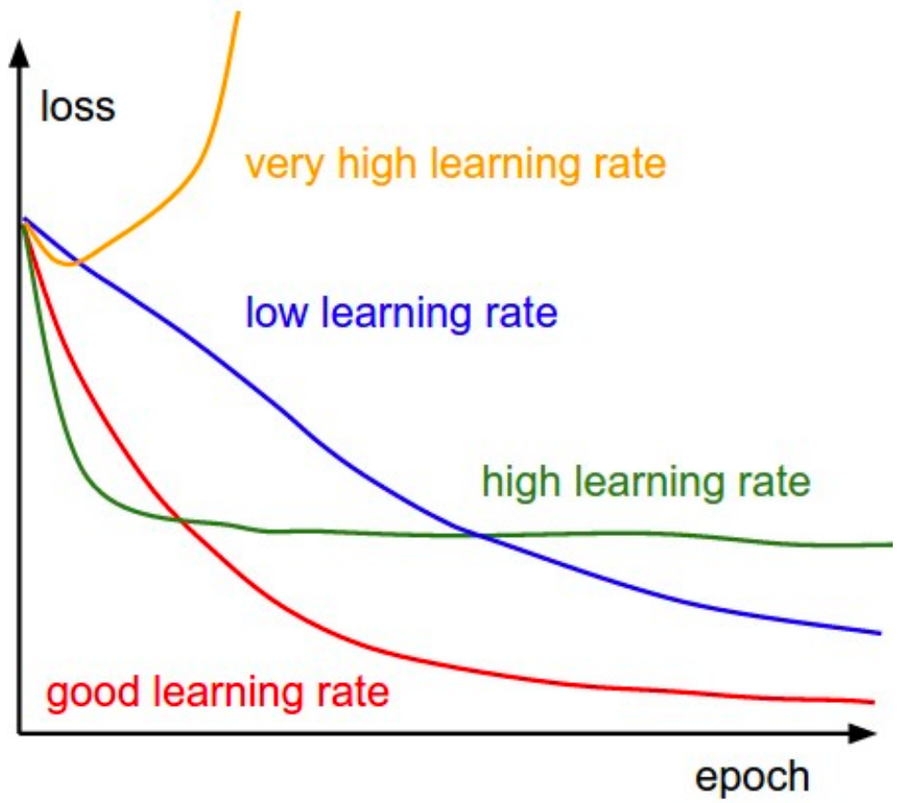
η_t : step size / learning rate



How to set the step size?

Theory answer: it depends …

Practice answer: manually tune



Gradient descent visualization credit: Sunil Jangir

Step size cartoon credit: Stanford CS 231N

How to Set the Gradient Descent Step Size?

Theory answer: it depends on the problem structure

non-smooth

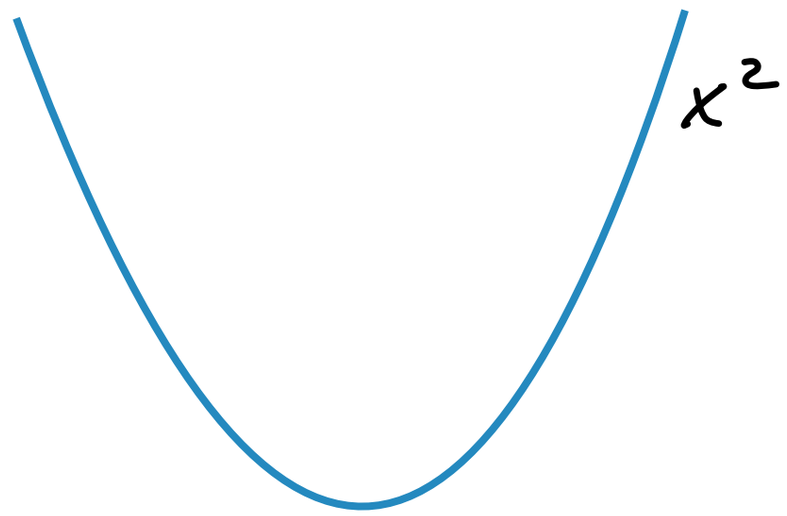
$$\|\nabla f(x)\| \le G$$

$$\eta_t = \frac{\|x_0 - x^*\|}{G\sqrt{T}}$$

$$T = \frac{G^2\|x_0 - x^*\|^2}{\epsilon^2}$$ optimal

smooth

$$\|\nabla f(x) - \nabla f(y)\| \le \beta\|x - y\|$$



$$\eta_t = 1/\beta$$

$$T = \frac{\beta\|x_0 - x^*\|^2}{\epsilon}$$

AGD: $$T = \frac{\beta\|x_0 - x^*\|^2}{\sqrt{\epsilon}}$$ optimal

How to Set the Gradient Descent Step Size?

Theory answer: it depends on the problem structure

Caveats:

- Step sizes depend on several parameters (smoothness, gradient norm, distance to x^*, …)
- Parameters are often unknown and hard to tune

The dream:

- Automatically learn the step size
- Adapt to (local or global) smoothness and convexity
- Universal algorithms that achieve optimal convergence in the smooth and non-smooth settings simultaneously

The Plan for this Lecture

- Introduce and analyze a variant of gradient descent that uses the gradients observed to set the step sizes
- We will show that the algorithm is universal: it automatically adapts to the problem structure (non-smooth or smooth)
- We will show that the algorithm adapts to the problem parameters G or β
- We will assume throughout that the domain has bounded diameter and we know an upper bound R on the diameter
- There exist algorithms that adapt to the diameter R of the domain, but we will not discuss them in these lectures

Adaptive Gradient Descent

$$\min_{x \in \mathbb{R}^d} f(x)$$

[Duchi, Hazan, Singer; McMahan and Streeter 2010]

Adagrad (scalar version)

$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

$$\eta_t = \frac{R}{\sqrt{\sum_{s=1}^{t} \|\nabla f(x_s)\|^2}}$$

only needed to get the optimal dependence on R (in practice we use $\frac{1}{\sqrt{\dots}}$)

Adaptive Gradient Descent

$$\min_{x \in K} f(x)$$

[Duchi, Hazan, Singer; McMahan and Streeter 2010]

Adagrad (scalar version)

$$x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \| u - x_t \|^2 \right\}$$

$$\eta_t = \frac{R}{\sqrt{\sum_{s=1}^{t} \|\nabla f(x_s)\|^2}}$$

Same as projected GD

$$x_{t+1} = \Pi_K \left(x_t - \eta_t \nabla f(x_t) \right)$$

Adaptive Gradient Descent

$$\min_{x \in K} f(x)$$

[Duchi, Hazan, Singer; McMahan and Streeter 2010]

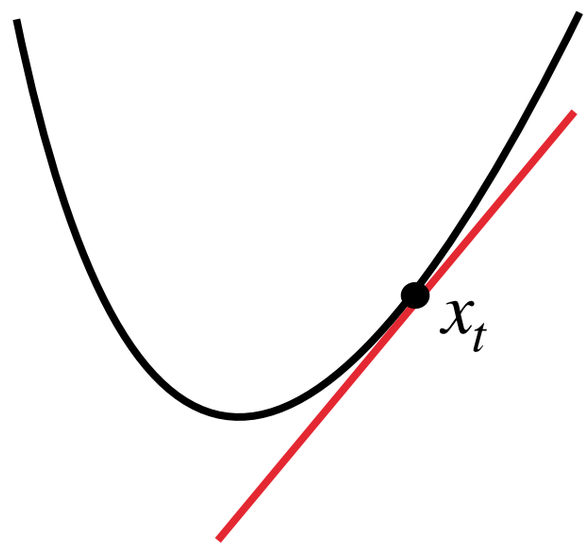
Adagrad (scalar version)

$$x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \| u - x_t \|^2 \right\}$$

$$\eta_t = \frac{R}{\sqrt{\sum_{s=1}^{t} \|\nabla f(x_s)\|^2}}$$

Gradient descent algorithm:

$$x_{t+1} = \arg\min_{u \in K} \left\{ \underbrace{f(x_t) + \langle \nabla f(x_t), u - x_t \rangle}_{\text{linear approx}} + \frac{1}{\eta_t} \cdot \underbrace{\frac{1}{2}\|u - x_t\|^2}_{\text{movement}} \right\}$$



Adaptive Gradient Descent

$$\min_{x \in K} f(x)$$

[Duchi, Hazan, Singer; McMahan and Streeter 2010]

Adagrad (scalar version)

$$x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \| u - x_t \|^2 \right\}$$

$$\eta_t = \frac{R}{\sqrt{\sum_{s=1}^{t} \|\nabla f(x_s)\|^2}}$$

A note on the feasible domain K: $\nabla f(x^*) = 0 \quad x^* \in K$

- The step choice above is suitable for domains that are essentially unconstrained: K contains a global minimum
- We will discuss extensions to general domains at the end of the lecture

Adaptive Gradient Descent

$$\min_{x \in K} f(x)$$

Adagrad (scalar version)

Let $x_1 \in K, R \geq \max_{x,y \in K} \|x - y\|$

For $t = 1, \ldots, T$:

$$\eta_t = \frac{R}{\sqrt{\sum_{s=1}^{t} \|\nabla f(x_s)\|^2}}$$

$$x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \|u - x_t\|^2 \right\}$$

Return $\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$

Adaptive Gradient Descent

$$\min_{x \in K} f(x)$$

Adagrad (scalar version)

Let $x_1 \in K, R \geq \max_{x,y \in K} \|x - y\|$

For $t = 1, \ldots, T$:

Assumption: K contains global min $x^* \in K$ and $\nabla f(x^*) = 0$

$$\eta_t = \frac{R}{\sqrt{\sum_{s=1}^{t} \|\nabla f(x_s)\|^2}}$$

$$x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \|u - x_t\|^2 \right\}$$

Return $\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$

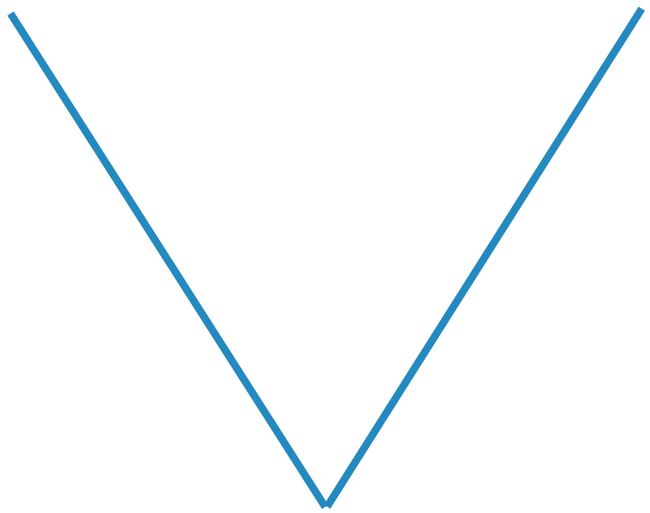
The Unreasonable Effectiveness of Adagrad

AutoML

It automatically adapts to problem structure

non-smooth

$$\|\nabla f(x)\| \leq G$$



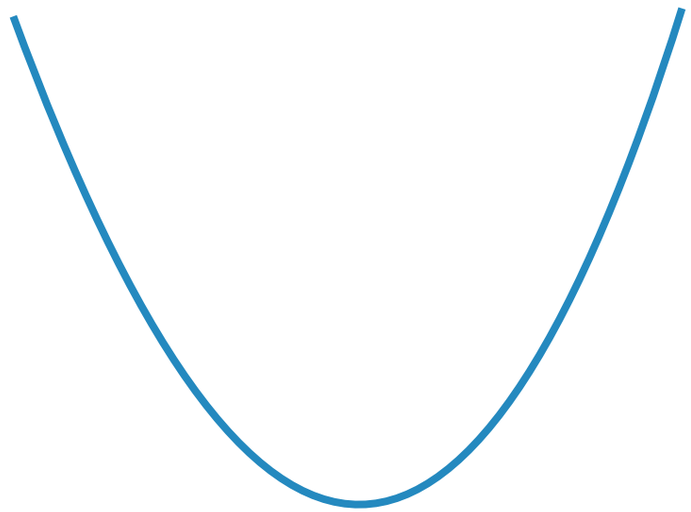
$$T = \frac{G^2 R^2}{\epsilon^2}$$

optimal

[Duchi et al., McMahan & Streeter 2010]

smooth

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$



$$T = \frac{\beta R^2}{\epsilon}$$

non-accelerated smooth rate

[Levy 2017, Levy et al. 2018]

[E., Nguyen, Vladu 2020]

$$R = \max_{t \in [T]} \|x_t - x^*\|$$

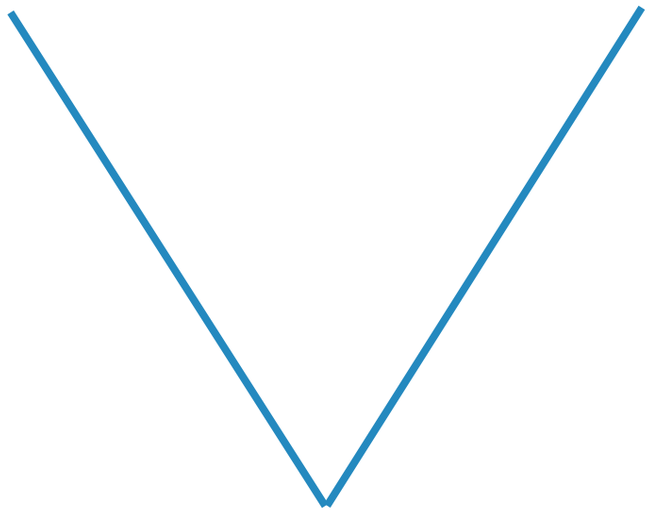
The Unreasonable Effectiveness of Adagrad

AutoML

It automatically adapts to problem structure

non-smooth

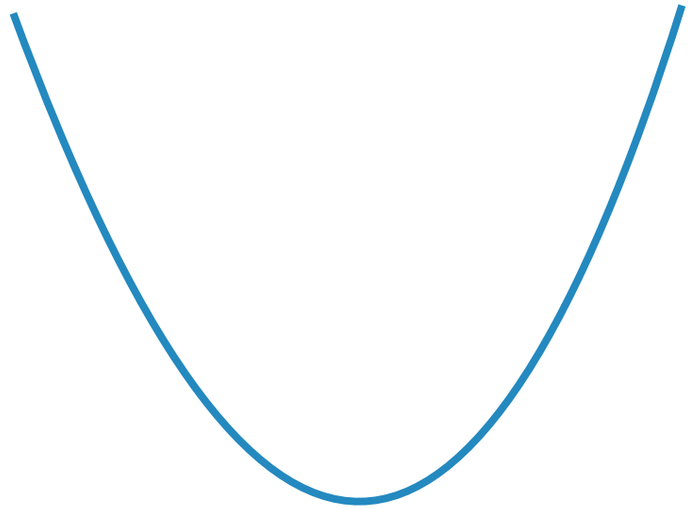
$$\|\nabla f(x)\| \leq G$$



$$\|x\| = \|x\|_2 = \sqrt{\sum_{i=1}^{} x_i^2}$$

smooth

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\|$$



Intuition

$\|\nabla f(x_t)\|^2$ stays constant

$$\eta_t = \left(\sum_{s=1}^{t} \|\nabla f(x_s)\|^2\right)^{-1/2} = O\left(\frac{1}{\sqrt{t}}\right)$$

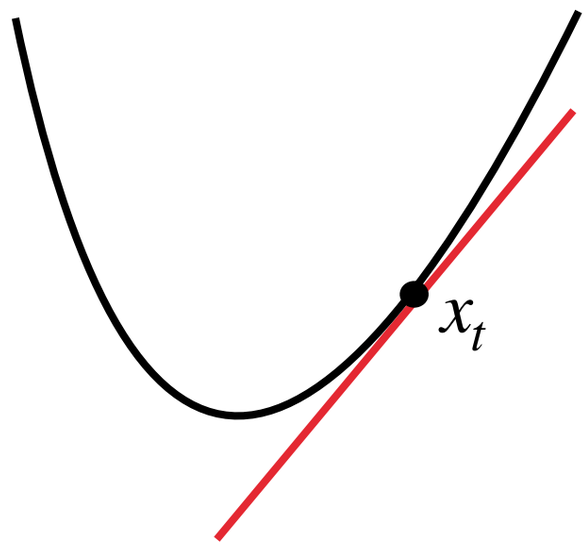
$\|\nabla f(x_t)\|$ decays at a $\frac{1}{t}$ rate

$$\eta_t = \left(\sum_{s=1}^{t} \|\nabla f(x_s)\|^2\right)^{-1/2} = O(1)$$

Adagrad Analysis

- The analysis combines the non-adaptive gradient descent analysis with some additional results
- We can use the potential-based analysis from the primer lecture
- We will illustrate a different (but related) analysis that leverages the optimization perspective of the GD update

$$x_{t+1} = \arg\min_{u \in K} \left\{ \underbrace{f(x_t) + \langle \nabla f(x_t), u - x_t \rangle}_{\text{linear approx}} + \frac{1}{\eta_t} \cdot \underbrace{\frac{1}{2}\|u - x_t\|^2}_{\text{movement}} \right\}$$



Adagrad Analysis: Non-smooth

As before, we assume that the gradients are bounded:

$$\|\nabla f(x)\| \leq G \quad \forall x \in K$$

E.g : when f is G-Lipschitz : $|f(x)-f(y)| \leq G\|x-y\|$

Adagrad Analysis: Non-smooth

As before, we assume that the gradients are bounded:

$$\|\nabla f(x)\| \leq G \quad \forall x \in K$$

Algorithm returns the uniform average:

$$\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$$

Adagrad Analysis: Non-smooth

As before, we assume that the gradients are bounded:

$$\|\nabla f(x)\| \leq G \quad \forall x \in K$$

Algorithm returns the uniform average:

$$\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$$

By convexity:

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right)$$

Adagrad Analysis: Non-smooth

As before, we assume that the gradients are bounded:

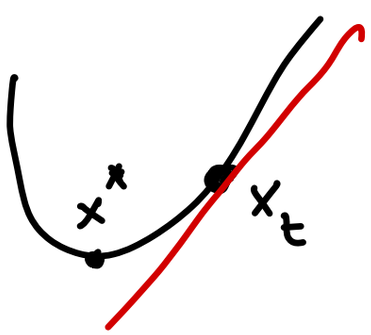
$$\|\nabla f(x)\| \leq G \quad \forall x \in K$$

Algorithm returns the uniform average:

$$\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$$

By convexity:

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) \quad \text{usual defn. of convexity}$$



$$\leq \frac{1}{T}\sum_{t=1}^{T}\left\langle \nabla f(x_t), x_t - x^* \right\rangle \quad \text{first-order characterization}$$

$$f(x^*) \geq f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle$$

Adagrad Analysis: Non-smooth

Next, we upper bound the inner product terms $\langle \nabla f(x_t), x_t - x^* \rangle$

Recall the update rule:

$$x_{t+1} = \arg\min_{u \in K} \underbrace{\left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \|u - x_t\|^2 \right\}}_{\phi(u)}$$

Recall the first-order optimality condition for $u^* = \arg\min_{u \in K} \phi(u)$:

$K = \mathbb{R}^d$: $\nabla \phi(u^*) = 0$

K general : $\langle \nabla \phi(u^*), u^* - u \rangle \leq 0 \quad \forall\, u \in K$

$$\Rightarrow \left\langle \nabla f(x_t) + \frac{1}{\eta_t}(x_{t+1} - x_t),\ x_{t+1} - x^* \right\rangle \leq 0$$

Adagrad Analysis: Non-smooth

Next, we upper bound the inner product terms $\langle \nabla f(x_t), x_t - x^* \rangle$

Recall the update rule:

$$x_{t+1} = \arg\min_{u \in K} \left\{ \left\langle \nabla f(x_t), u - x_t \right\rangle + \frac{1}{2\eta_t} \|u - x_t\|^2 \right\}$$

The optimality condition gives us:

$$\left\langle \nabla f(x_t), x_{t+1} - x^* \right\rangle \leq \frac{1}{\eta_t} \left\langle x_t - x_{t+1}, x_{t+1} - x^* \right\rangle$$ rearrangement

$$= \frac{1}{2\eta_t} \left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 - \|x_{t+1} - x_t\|^2 \right)$$

$$ab = \frac{1}{2}(a+b)^2 - \frac{1}{2}a^2 - \frac{1}{2}b^2$$

Adagrad Analysis: Non-smooth

Next, we upper bound the inner product terms $\langle \nabla f(x_t), x_t - x^* \rangle$

Recall the update rule:

$$x_{t+1} = \arg\min_{u \in K} \left\{ \left\langle \nabla f(x_t), u - x_t \right\rangle + \frac{1}{2\eta_t} \|u - x_t\|^2 \right\}$$

The optimality condition gives us:

$$\left\langle \nabla f(x_t), x_{t+1} - x^* \right\rangle \leq \frac{1}{\eta_t} \left\langle x_t - x_{t+1}, x_{t+1} - x^* \right\rangle$$

$$= \frac{1}{2\eta_t} \left(\| x_t - x^* \|^2 - \| x_{t+1} - x^* \|^2 - \| x_t - x_{t+1} \|^2 \right)$$

Adagrad Analysis: Non-smooth

We have shown:

$$\langle \nabla f(x_t), x_{t+1} - x^* \rangle \leq \frac{1}{2\eta_t} \left(\| x_t - x^* \|^2 - \| x_{t+1} - x^* \|^2 - \| x_t - x_{t+1} \|^2 \right)$$

But we wanted to bound $\langle \nabla f(x_t), x_t - x^* \rangle \ldots$

Adagrad Analysis: Non-smooth

We have shown:

$$\left\langle \nabla f(x_t), x_{t+1} - x^* \right\rangle \leq \frac{1}{2\eta_t} \left(\| x_t - x^* \|^2 - \| x_{t+1} - x^* \|^2 - \| x_t - x_{t+1} \|^2 \right)$$

But we wanted to bound $\left\langle \nabla f(x_t), x_t - x^* \right\rangle \ldots$

To address this discrepancy, we write

$$\left\langle \nabla f(x_t), x_t - x^* \right\rangle = \left\langle \nabla f(x_t), x_{t+1} - x^* \right\rangle + \left\langle \nabla f(x_t), x_{t+1} - x_t \right\rangle$$

Cauchy-Schwartz: $\leq \|\nabla f(x_t)\| \cdot \|x_{t+1} - x_t\|$

gain will help us offset the loss

Adagrad Analysis: Non-smooth

We have shown:

$$\left\langle \nabla f(x_t), x_{t+1} - x^* \right\rangle \leq \frac{1}{2\eta_t} \left(\left\| x_t - x^* \right\|^2 - \left\| x_{t+1} - x^* \right\|^2 - \left\| x_t - x_{t+1} \right\|^2 \right)$$

But we wanted to bound $\left\langle \nabla f(x_t), x_t - x^* \right\rangle \ldots$

To address this discrepancy, we write

$$\begin{aligned}
\left\langle \nabla f(x_t), x_t - x^* \right\rangle &= \left\langle \nabla f(x_t), x_{t+1} - x^* \right\rangle + \left\langle \nabla f(x_t), x_{t+1} - x_t \right\rangle \\
&\leq \frac{1}{2\eta_t} \left\| x_t - x^* \right\|^2 - \frac{1}{2\eta_t} \left\| x_{t+1} - x^* \right\|^2 + \left\langle \nabla f(x_t), x_{t+1} - x_t \right\rangle - \frac{1}{2\eta_t} \left\| x_t - x_{t+1} \right\|^2
\end{aligned}$$

Adagrad Analysis: Non-smooth

We have shown:

$$\langle \nabla f(x_t), x_{t+1} - x^* \rangle \leq \frac{1}{2\eta_t} \left(\| x_t - x^* \|^2 - \| x_{t+1} - x^* \|^2 - \| x_t - x_{t+1} \|^2 \right)$$

But we wanted to bound $\langle \nabla f(x_t), x_t - x^* \rangle \ldots$

To address this discrepancy, we write

$$\begin{aligned}
\langle \nabla f(x_t), x_t - x^* \rangle &= \langle \nabla f(x_t), x_{t+1} - x^* \rangle + \langle \nabla f(x_t), x_{t+1} - x_t \rangle \\
&\leq \frac{1}{2\eta_t} \| x_t - x^* \|^2 - \frac{1}{2\eta_t} \| x_{t+1} - x^* \|^2 + \langle \nabla f(x_t), x_{t+1} - x_t \rangle - \frac{1}{2\eta_t} \| x_t - x_{t+1} \|^2
\end{aligned}$$

By Cauchy-Schwartz and the inequality $ab \leq \frac{\lambda}{2} a^2 + \frac{1}{2\lambda} b^2$: ($\lambda > 0$)

$$\begin{aligned}
\langle \nabla f(x_t), x_{t+1} - x_t \rangle - \frac{1}{2\eta_t} \| x_t - x_{t+1} \|^2 &\leq \| \nabla f(x_t) \| \| x_{t+1} - x_t \| - \frac{1}{2\eta_t} \| x_t - x_{t+1} \|^2 \\
&\leq \frac{\eta_t}{2} \| \nabla f(x_t) \|^2 \qquad \lambda = \eta_t
\end{aligned}$$

Adagrad Analysis: Non-smooth

We have shown:

$$\left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq \frac{1}{2\eta_t} \left\| x_t - x^* \right\|^2 - \frac{1}{2\eta_t} \left\| x_{t+1} - x^* \right\|^2 + \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Adagrad Analysis: Non-smooth

We have shown:

$$\left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq \frac{1}{2\eta_t} \left\| x_t - x^* \right\|^2 - \frac{1}{2\eta_t} \left\| x_{t+1} - x^* \right\|^2 + \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Summing up and collecting terms:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq \sum_{t=2}^{T} \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \underbrace{\left\| x_t - x^* \right\|^2}_{\leq R^2} + \frac{1}{2\eta_1} \underbrace{\left\| x_2 - x^* \right\|^2}_{\leq R^2} + \sum_{t=1}^{T} \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Adagrad Analysis: Non-smooth

We have shown:

$$\left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq \frac{1}{2\eta_t} \left\| x_t - x^* \right\|^2 - \frac{1}{2\eta_t} \left\| x_{t+1} - x^* \right\|^2 + \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Summing up and collecting terms:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq \sum_{t=2}^{T} \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \underbrace{\left\| x_t - x^* \right\|^2}_{\leq R^2} + \frac{1}{2\eta_1} \underbrace{\left\| x_2 - x^* \right\|^2}_{\leq R^2} + \sum_{t=1}^{T} \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

$$\leq \frac{R^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Adagrad Analysis: Non-smooth

We have shown:

$$\left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq \frac{1}{2\eta_t} \left\| x_t - x^* \right\|^2 - \frac{1}{2\eta_t} \left\| x_{t+1} - x^* \right\|^2 + \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Summing up and collecting terms:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq \sum_{t=2}^{T} \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right) \underbrace{\left\| x_t - x^* \right\|^2}_{\leq R^2} + \frac{1}{2\eta_1} \underbrace{\left\| x_2 - x^* \right\|^2}_{\leq R^2} + \sum_{t=1}^{T} \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

$$\leq \frac{R^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Note: we crucially relied on our assumption that K has bounded diameter in order to telescope the sums.

Adagrad Analysis: Non-smooth

We have shown:

$$\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle \leq \frac{R^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Recall our choice of step sizes:

$$\eta_t = \frac{R}{\sqrt{\sum_{i=1}^{t} \|\nabla f(x_i)\|^2}}$$

Adagrad Analysis: Non-smooth

We have shown:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq \frac{R^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Recall the update rule for the step sizes:

$$\frac{R^2}{\eta_T} = R\sqrt{\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2} \qquad \sum_{t=1}^{T} \eta_t \left\| \nabla f(x_t) \right\|^2 = R \sum_{t=1}^{T} \frac{\left\| \nabla f(x_t) \right\|^2}{\sqrt{\sum_{i=1}^{t} \left\| \nabla f(x_i) \right\|^2}}$$

Adagrad Analysis: Non-smooth

We have shown:

$$\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle \leq \frac{R^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Recall the update rule for the step sizes:

$$\frac{R^2}{\eta_T} = R\sqrt{\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2} \qquad \sum_{t=1}^{T} \eta_t \left\| \nabla f(x_t) \right\|^2 = R \sum_{t=1}^{T} \frac{\left\| \nabla f(x_t) \right\|^2}{\sqrt{\sum_{i=1}^{t} \left\| \nabla f(x_i) \right\|^2}}$$

Lemma: For any positive numbers $a_1, \ldots, a_T$, we have

$$\sqrt{\sum_{t=1}^{T} a_t} \leq \sum_{t=1}^{T} \frac{a_t}{\sqrt{\sum_{s=1}^{t} a_s}} \leq 2\sqrt{\sum_{t=1}^{T} a_t}$$

Proof idea: think of $\frac{a_t}{\sqrt{\sum_{s=1}^{t} a_s}}$ as $\frac{dx}{\sqrt{x}}$ and recall that $\int \frac{dx}{\sqrt{x}} = \sqrt{x}$

Adagrad Analysis: Non-smooth

We have shown:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq \frac{R^2}{2\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \left\| \nabla f(x_t) \right\|^2$$

Recall the update rule for the step sizes:

$$\frac{R^2}{\eta_T} = R\sqrt{\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2} \qquad \sum_{t=1}^{T} \eta_t \left\| \nabla f(x_t) \right\|^2 = R \sum_{t=1}^{T} \frac{\left\| \nabla f(x_t) \right\|^2}{\sqrt{\sum_{i=1}^{t} \left\| \nabla f(x_i) \right\|^2}}$$

Using the inequality, we obtain

$$\sum_{t=1}^{T} \frac{\left\| \nabla f(x_t) \right\|^2}{\sqrt{\sum_{i=1}^{t} \left\| \nabla f(x_i) \right\|^2}} \leq 2\sqrt{\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2}$$

Adagrad Analysis: Non-smooth

We have shown:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq 3R \sqrt{\sum_{t=1}^{T} \underbrace{\left\| \nabla f(x_t) \right\|^2}_{\leq G^2}} \qquad \leq 3RG\sqrt{T}$$

Adagrad Analysis: Non-smooth

We have shown:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq 3R \sqrt{\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2}$$

Finally, we use the bounded gradient assumption:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq 3RG\sqrt{T}$$

Adagrad Analysis: Non-smooth

We have shown:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq 3R \sqrt{\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2}$$

Finally, we use the bounded gradient assumption:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq 3RG\sqrt{T}$$

We have thus obtained our convergence guarantee:

$$f(\overline{x}_T) - f(x^*) \leq \frac{1}{T} \sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle$$

$$\leq O\left(\frac{RG}{\sqrt{T}}\right)$$

Adagrad Analysis: Smooth

We assume f is β-smooth:

$$\| \nabla f(x) - \nabla f(y) \| \leq \beta \| x - y \| \qquad \forall x, y$$

Adagrad Analysis: Smooth

We assume f is β-smooth:

$$\| \nabla f(x) - \nabla f(y) \| \leq \beta \| x - y \| \qquad \forall x, y$$

Smoothness implies (exercise):

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\beta} \| \nabla f(y) - \nabla f(x) \|^2 \qquad \forall x, y$$

convexity

extra gain (from smoothness)

Adagrad Analysis: Smooth

We assume f is β-smooth:

$$\| \nabla f(x) - \nabla f(y) \| \leq \beta \| x - y \| \qquad \forall x, y$$

Smoothness implies (exercise):

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\beta} \| \nabla f(y) - \nabla f(x) \|^2 \qquad \forall x, y$$

Setting $y = x_t$ and $x = x^*$, and using that $\nabla f(x^*) = 0$:

$$f(x_t) - f(x^*) \leq \langle \nabla f(x_t), x_t - x^* \rangle - \frac{1}{2\beta} \| \nabla f(x_t) \|^2$$

Adagrad Analysis: Smooth

We assume f is β-smooth:

$$\| \nabla f(x) - \nabla f(y) \| \leq \beta \| x - y \| \qquad \forall x, y$$

Smoothness implies (exercise):

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\beta} \| \nabla f(y) - \nabla f(x) \|^2 \qquad \forall x, y$$

Setting $y = x_t$ and $x = x^*$, and using that $\nabla f(x^*) = 0$:

$$f(x_t) - f(x^*) \leq \langle \nabla f(x_t), x_t - x^* \rangle - \frac{1}{2\beta} \| \nabla f(x_t) \|^2$$

Thus we have

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T} \sum_{t=1}^{T} \left(f(x_t) - f(x^*) \right)$$

$$\leq \frac{1}{T} \left(\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2 \right)$$

Adagrad Analysis: Smooth

We have:

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T}\left(\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2 \right)$$

Adagrad Analysis: Smooth

We have:

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T}\left(\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2 \right)$$

We have shown:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq 3R \sqrt{\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2}$$

Adagrad Analysis: Smooth

We have:

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T}\left(\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2 \right)$$

We have shown:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq 3R \sqrt{\sum_{t=1}^{T} \| \nabla f(x_t) \|^2}$$

The gain offsets the loss:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2 \leq \frac{3}{2} R \sqrt{\sum_{t=1}^{T} \| \nabla f(x_t) \|^2} - \frac{1}{2\beta} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2$$

t $\qquad$ t^2

t^2 grows faster than t: eventually, gain will exceed loss

Adagrad Analysis: Smooth

We have:

$$f(\bar{x}_T) - f(x^*) \le \frac{1}{T}\left(\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2 \right)$$

We have shown:

$$\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle \le 3R \sqrt{\sum_{t=1}^{T} \| \nabla f(x_t) \|^2}$$

The gain offsets the loss:

$$\sum_{t=1}^{T} \langle \nabla f(x_t), x_t - x^* \rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2 \le \frac{3}{2} R \sqrt{\sum_{t=1}^{T} \| \nabla f(x_t) \|^2} - \frac{1}{2\beta} \sum_{t=1}^{T} \| \nabla f(x_t) \|^2$$

$$\le \max_{z \ge 0} \left\{ \frac{3}{2} R z - \frac{1}{2\beta} z^2 \right\} = O(\beta R^2)$$

$\phi(z) = az - bz^2$: concave function of z

$a, b > 0$ $\quad \phi'(z) = 0 \Rightarrow z^* = \frac{a}{2b}$ and $\phi(z^*) = \frac{a^2}{4b}$

Adagrad Analysis: Smooth

We have:

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T}\left(\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2 \right)$$

We have shown:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq 3R \sqrt{\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2}$$

The gain offsets the loss:

$$\begin{aligned}
\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2 &\leq \frac{3}{2} R \sqrt{\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2} - \frac{1}{2\beta} \sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2 \\
&\leq \max_{z \geq 0} \left\{ \frac{3}{2} R z - \frac{1}{2\beta} z^2 \right\} \\
&= O\left(\beta R^2\right)
\end{aligned}$$

Adagrad Analysis: Smooth

We have:

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T}\left(\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2 \right)$$

We have shown:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle \leq 3R \sqrt{\sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2}$$

The gain offsets the loss:

$$\sum_{t=1}^{T} \left\langle \nabla f(x_t), x_t - x^* \right\rangle - \frac{1}{2\beta} \sum_{t=1}^{T} \left\| \nabla f(x_t) \right\|^2 \leq O\left(\beta R^2\right)$$

Thus we have our final convergence guarantee:

$$f(\bar{x}_T) - f(x^*) \leq O\left(\frac{\beta R^2}{T}\right)$$

Adagrad for Constrained Optimization

$$\min_{x \in K} f(x)$$

unconstrained: $\nabla f(x^*) = 0$

constrained: $\nabla f(x^*) \neq 0$

Intuition: as we approach x^*, the gradient does not decrease but the iterate movement $\|x_{t+1} - x_t\|$ does

Adagrad+ algorithm:

$$x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \| u - x_t \|^2 \right\}$$

$$\frac{1}{\eta_{t+1}^2} = \frac{1}{\eta_t^2} \left(1 + \| x_{t+1} - x_t \|^2 \right)$$

Adagrad for Constrained Optimization

$$\min_{x \in K} f(x)$$

unconstrained: $\nabla f(x^*) = 0$

constrained: $\nabla f(x^*) \neq 0$

Intuition: as we approach x^*, the gradient does not decrease but the iterate movement $\|x_{t+1} - x_t\|$ does

Adagrad+ algorithm:

$$x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \| u - x_t \|^2 \right\}$$

$$\frac{1}{\eta_{t+1}^2} = \frac{1}{\eta_t^2} \left(1 + \frac{\| x_{t+1} - x_t \|^2}{R^2} \right)$$

Good to scale the updates so that the steps change by at most a constant factor

Adagrad for Constrained Optimization

$$\min_{x \in K} f(x)$$

unconstrained: $\nabla f(x^*) = 0$

constrained: $\nabla f(x^*) \neq 0$

Intuition: as we approach x^*, the gradient does not decrease but the iterate movement $\|x_{t+1} - x_t\|$ does

Adagrad+ algorithm:

$$x_{t+1} = \arg\min_{u \in K} \left\{ \langle \nabla f(x_t), u - x_t \rangle + \frac{1}{2\eta_t} \| u - x_t \|^2 \right\}$$

$$\frac{1}{\eta_{t+1}^2} = \frac{1}{\eta_t^2} \left(1 + \frac{\| x_{t+1} - x_t \|^2}{R^2} \right)$$

$$\eta_t = \frac{R}{\sqrt{R^2 \frac{1}{\eta_1^2} + \sum_{i=1}^{t-1} \frac{1}{\eta_i^2} \| x_{i+1} - x_i \|^2}}$$

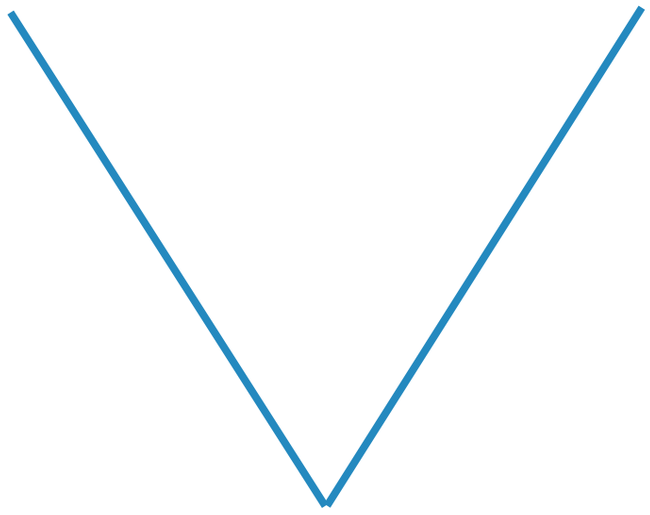
$K = \mathbb{R}^d$: $\frac{1}{\eta_i^2} \| x_{i+1} - x_i \|^2 = \| \nabla f(x_i) \|^2$

Adagrad+ for Constrained Optimization

It automatically adapts to problem structure

non-smooth

$$\|\nabla f(x)\| \leq G$$



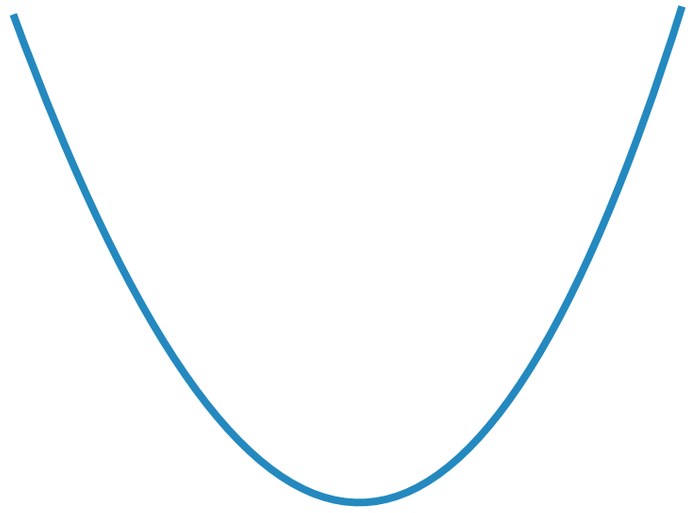
$$T = \tilde{O}\left(\frac{1}{\epsilon^2}\right)$$

nearly-optimal (up to logs)

[E., Nguyen, Vladu 2020]

smooth

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\|$$



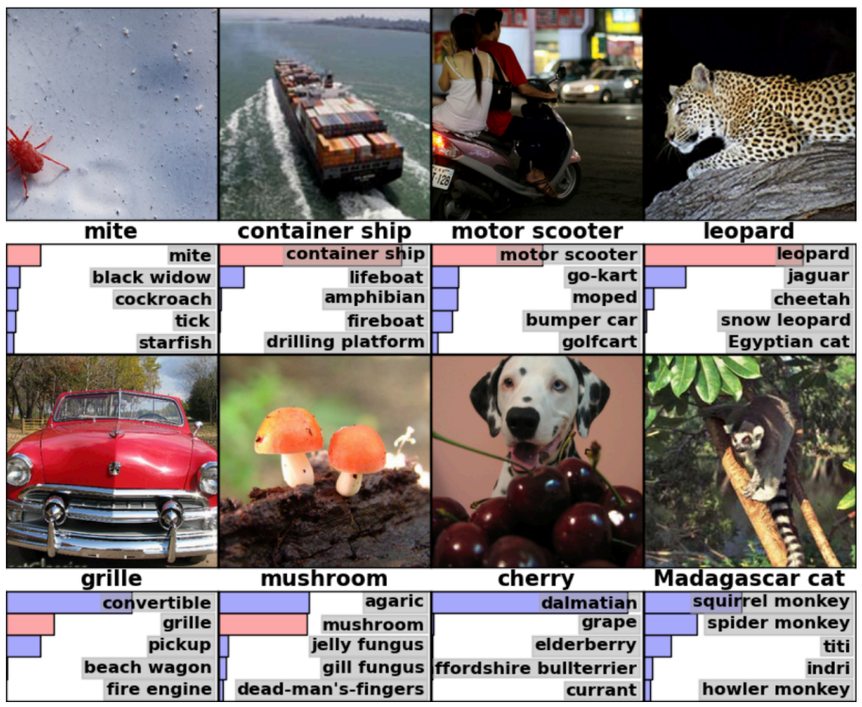
$$T = O\left(\frac{1}{\epsilon}\right)$$

non-accelerated smooth rate

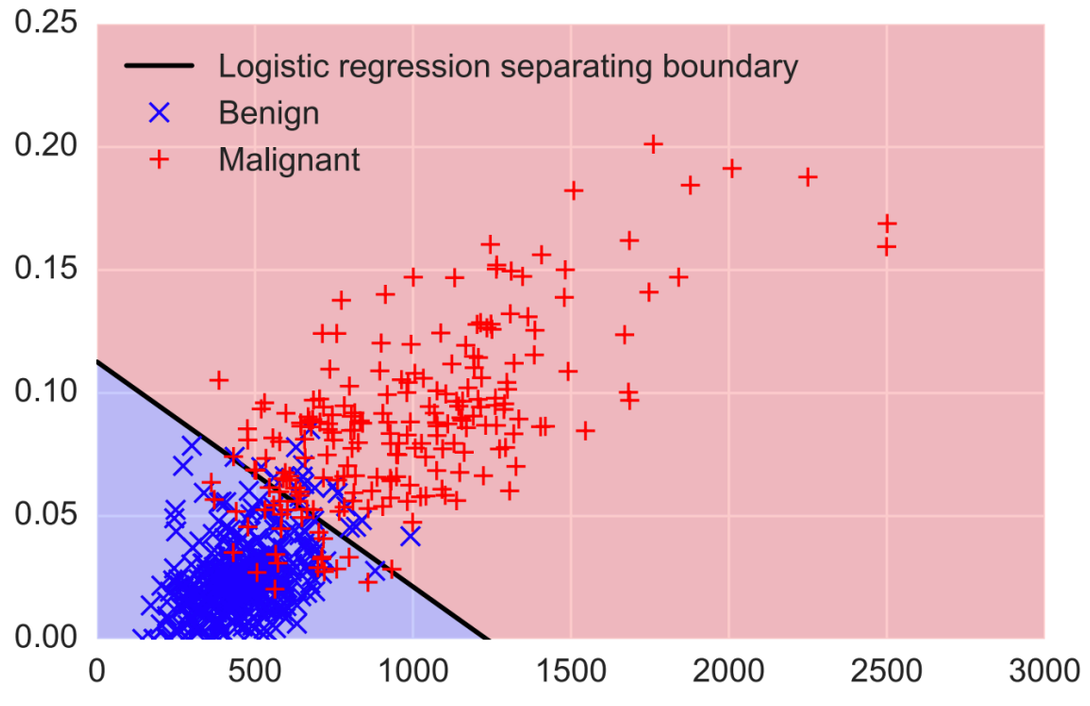
[E., Nguyen, Vladu 2020]

Machine Learning Applications

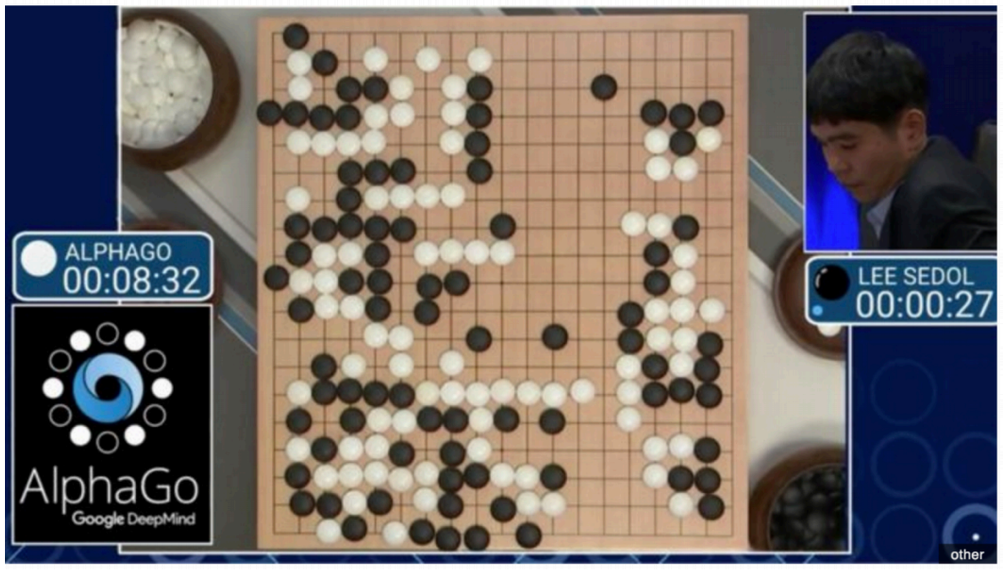
ImageNet Classification



Cancer Classification



AlphaGo



Power Demand Regression

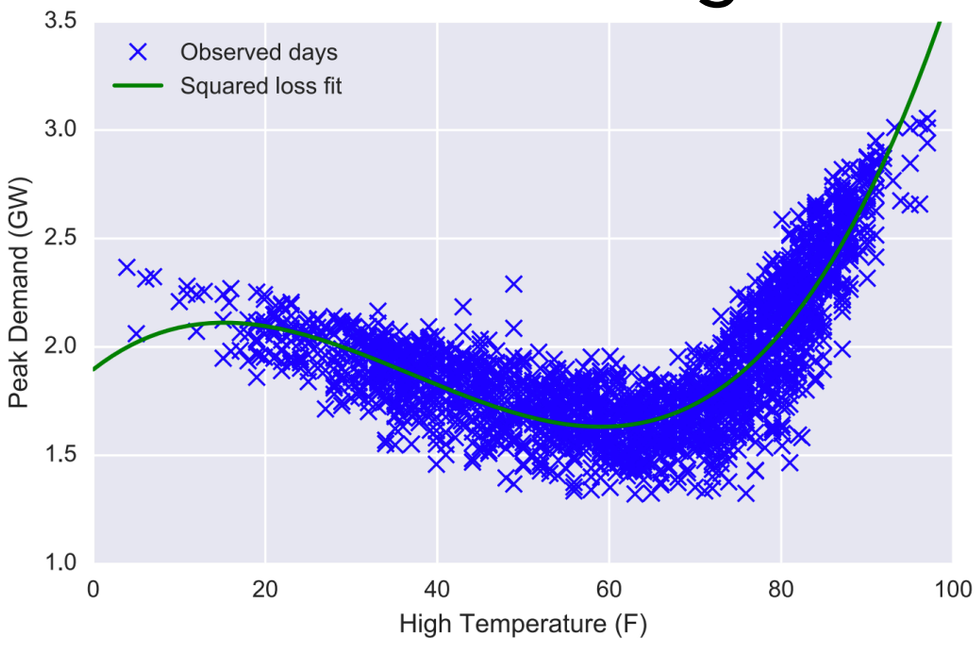


Image credits: Zico Kolter

Adagrad Family

- **Adagrad** (Adaptive Gradient)
  [Duchi et al., McMahan and Streeter, 2010]
- **Adadelta** [Zeiler, 2012]
- **RMSProp** [Hinton, 2014]
- **Adam** (Adaptive Moment Estimation)
  [Kingma and Ba, 2015]
- **AdaMax** [Kingma and Ba, 2015]
- **Nadam** (Nesterov-accelerated Adaptive Moment Estimation)
  [Dozat, 2016]

Adagrad Family

- **Adagrad** (Adaptive Gradient)
  [Duchi et al., McMahan and Streeter, 2010]

7067 citations

- **Adadelta** [Zeiler, 2012]
- **RMSProp** [Hinton, 2014]

- **Adam** (Adaptive Moment Estimation)
  [Kingma and Ba, 2015]

53803 citations

- **AdaMax** [Kingma and Ba, 2015]
- **Nadam** (Nesterov-accelerated Adaptive Moment Estimation)
  [Dozat, 2016]