The Five Miracles of Mirror Descent

Hadar Tal

hadar.tal@mail.huji.ac.il

This paper is a summary of the educational materials and lectures from

- **The Five Miracles of Mirror Descent** by Professor Sebastien Bubeck, Microsoft Research (Claire Boyer's notes)
- **Wikipedia**
- **3Blue1Brown** YouTube channel

Winter 2024

Contents

1 The First Miracle: Robustness **1**
1.1 Gradient Descent 1
  1.1.1 Analysis of the Gradient Descent Algorithm 1

2 The Second Miracle: Potential Based **3**
2.1 Experts Problem 3
    Approach 1: Gradient Descent 3
    Approach 2: Multiplicative Weights Update (MWU) 4
2.2 Mirror Descent 4

3 The Third Miracle: **7**
3.1 Coupon Collector Phenomenon 7
3.2 Metrical Task Systems (MTS) 7
  3.2.1 Deterministic Strategies on uniform metric 8
  3.2.2 Randomized Strategies for Uniform Metric 8
3.3 Beyond Uniform Spaces: weighted stars 9
  3.3.1 Wasserstein Distance 9
  3.3.2 Continues space problem 10

4 The Fourth Miracle: **11**

5 The Fifth Miracle: **13**

Chapter 1

The First Miracle: Robustness

Let f be a convex function, and let x^* be a minimizer of f.

1.1 Gradient Descent

Definition 1.1.1. *Gradient Descent*

$$x_{t+1} = x_t - \eta \nabla f(x_t) \tag{1.1}$$

It holds that:

$$f(x^*) \geq f(x_t) + \nabla f(x_t) \cdot (x^* - x_t) \tag{1.2}$$

$$0 \leq f(x_t) - f(x^*) \leq \nabla f(x_t) \cdot (x_t - x^*) \tag{1.3}$$

1.1.1 Analysis of the Gradient Descent Algorithm

$$\begin{aligned}
&\|a\|^2 = \|b\|^2 + \|a-b\|^2 \\
&\|b\|^2 = \|a\|^2 - \|a-b\|^2 = \|a\|^2 - (\|a\|^2 - 2a \cdot b + \|b\|^2) = 2a \cdot b - \|b\|^2
\end{aligned}$$

Then we have:

$$\begin{aligned}
\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2 &= -2\eta(x^* - x_t) \cdot \nabla f(x_t) - \eta^2 \|\nabla f(x_t)\|^2 \\
&= 2\eta(x_t - x^*) \cdot \nabla f(x_t) - \eta^2 \|\nabla f(x_t)\|^2 \\
&\geq 2(f(x_t) - f(x^*)) - \eta^2 L^2
\end{aligned}$$

Where the last inequality follows from the convexity and the Lipschitz continuity of f.

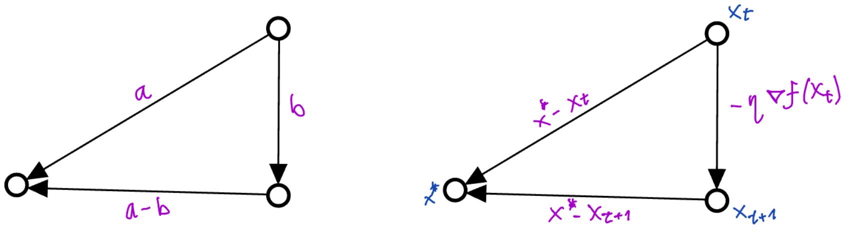

Figure 1.1: Gradient Descent

Then if we sum the above inequality from $t = 1$ to T, we get:

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

In fact, this is a specific case of the Fundamental Inequality of Optimization.

Theorem 1.1.1. *Fundamental Inequality of Optimization (unconstrained version)*
Suppose $x_{t+1} = x_t - \eta g_t$ *for all* t*, where* $g_1, \ldots, g_T \in \mathbb{R}^d$ *are arbitrary vectors. Then for all* $x^* \in \mathbb{R}^d$ *it holds that*

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

Proof. Fundamental Inequality of Optimization
The proof tracks $\|x_t - x^*\|^2$ as a “potential”. First write

$$\|x_{t+1} - x^*\|^2 = \|(x_t - x^*) - \eta g_t\|^2 = \|x_t - x^*\|^2 - 2\eta g_t \cdot (x_t - x^*) + \eta^2\|g_t\|^2,$$

that is,

$$\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 = 2\eta g_t \cdot (x_t - x^*) - \eta^2\|g_t\|^2.$$

Summing over $t = 1, \ldots, T$ and telescoping terms, we obtain

$$\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2 = 2\eta\sum_{t=1}^{T} g_t \cdot (x_t - x^*) - \eta^2\sum_{t=1}^{T}\|g_t\|^2.$$

Organizing terms, we conclude:

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

□

SGD , choosing eta

Chapter 2

The Second Miracle: Potential Based

2.1 Experts Problem

At each time step, the player picks an action $I_t \in [n]$ (we have n experts) and the adversary picks a loss vector $l_t \in 0, 1^n$. The player incurs loss $l_t(I_t)$ and the goal is to minimize the regret:

$$\text{Regret}_T(i) = \sum_{t=1}^{T} (l_t(I_t) - l_t(i)) \tag{2.1}$$

We consider the case where in each time step the player chooses an action from a distribution $\vec{p}$ over the n experts (a vector from the simplex):

$$\vec{p} \in \mathcal{K} := \triangle_n = \{\vec{p} \in \mathbb{R}^n_+ : p_i \geq 0, \sum_{i=1}^{n} p_i = 1\}$$

Approach 1: Gradient Descent

We can use gradient descent on $f_t(\vec{p_t}) = \vec{l_t} \cdot \vec{p}$, where $\vec{l_t}$ is the loss vector at time t. It holds that $\nabla f_t(\vec{p_t}) = \vec{l_t}$. We can use the analysis of the gradient descent algorithm for gradient descent of convex functions varying in time.
Let $q \in \triangle_n$ be any distribution. Then we have:

$$\begin{gathered} f_t(q) \geq f_t(\vec{p_t}) + \nabla f_t(q) \cdot (q - \vec{p_t}) \implies \\ f_t(\vec{p_t}) - f_t(q) \leq \nabla f_t(q) \cdot (\vec{p_t} - q) \end{gathered}$$

Then:

$$\begin{aligned} \|q - p_t\|^2 - \|q - p_{t+1}\|^2 &= -2\eta(q - p_t) \cdot \nabla f_t(p_t) - \eta^2 \|\nabla f_t(p_t)\|^2 \implies \\ f_t(\vec{p_t}) - f_t(q) &\leq \nabla f_t(q) \cdot (\vec{p_t} - q) = \frac{1}{2\eta}\left(\|q - \vec{p_t}\|^2 - \|q - \vec{p}_{t+1}\|^2\right) + \frac{\eta}{2}\|\nabla f_t(\vec{p_t})\|^2 \implies \\ \sum_{t=1}^{T} (f_t(\vec{p_t}) - f_t(q)) &\leq \frac{1}{2\eta}\left(\|q - \vec{p}_1\|^2 - \|q - \vec{p}_{T+1}\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T} \|\nabla f_t(\vec{p_t})\|^2 \\ &\leq \frac{1}{2\eta}\|q - \vec{p}_1\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \|\nabla f_t(\vec{p_t})\|^2 \\ &\leq \frac{1}{\eta} + \frac{\eta}{2} Tn = \mathbf{O}(\sqrt{Tn}) \end{aligned}$$

We have used the facts that:

- Both q and $\vec{p}_1$ are distributions, so $\|q - \vec{p}_1\|^2 \leq 2$.
- $\|\nabla f_t(\vec{p}_t)\|^2 \leq n$ (as the loss vector is in $0, 1^n$).

we can see that in this case, the rate of convergence DO depend on the dimension of the problem, in contrast to the non-varying case. The fact that the rate of convergence DO NOT depend on the dimension of the problem in GD is one of the reasons why GD is so useful in practice.

Approach 2: Multiplicative Weights Update (MWU)

The optimal convergence rate for the experts problem is $\sqrt{T \log n}$, which is achieved by the MWU algorithm. It is achieved by considering the sparsity of the OPT strategy (which is just a point mass on the best expert). We will modify GD such that it will move faster when far from a sparse point.

Approach 2: M

2.2 Mirror Descent

Endow $\mathcal{K}$ with a Riemannian structure:
$< \cdot, \cdot >_x$ for each $x \in \mathcal{K}$, the tangent space will be $\mathbb{R}^n$.

$$\text{Before:}$$
$$x_{t+1} = x_t - \eta \nabla f(x_t) \rightarrow f(x + dx) \approx f(x) + \langle \nabla f(x), dx \rangle$$
$$\text{Now:}$$
$$f(x + dx) \approx f(x) + \langle grad_M f(x), dx \rangle_x$$

We will consider M such that $\langle \cdot, \cdot \rangle_x = \nabla^2 \phi[\cdot, \cdot]$, where ϕ is a fixed convex function. The function ϕ encodes the geometry of the space, and this is where the prior information is encoded. We get that:

$$(dx)^T \nabla f(x) = \langle grad_M f(x), dx \rangle_x = dx^T \nabla^2 \phi(x) grad_M f(x) \Rightarrow$$
$$grad_M f(x) = [\nabla^2 \phi(x)]^{-1} \nabla f(x)$$

Idea: $x_{t+1} = x_t - \eta [\nabla^2 \phi(x_t)]^{-1} \nabla g_t$
Question: what is the potential? (how do we analyze this algorithm?)
Miracle: the potential is the Bregman divergence of ϕ ($D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle_x$) [Nemiroski, 1979]

Algorithm 1: Continues time Mirror Descent

$g : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ is a smooth vector field

$$\frac{d}{dt} x_t = -[\nabla^2 \phi(x_t)]^{-1} (\eta g(t) + \lambda(t))$$

Where $\lambda(t) \in N_{\mathcal{K}}(x_t), x(t) \in \mathcal{K}$ is a lagrange multiplier - it is an element of the normal cone of $\mathcal{K}$ at x_t.
This is the same as the following:

$$x(t + dt) = \operatorname*{argmin}_{x \in \mathcal{K}} \left\{ \eta g(t) \cdot x + \frac{1}{2} \|x - x(t)\|^2_{x(t)} \right\}$$

where $\|h\|_x^2 = \nabla^2 \phi(x)[h, h]$ is the norm induced by the Riemannian metric.

It is the same as the proximal view of GD, except that the proximal operator is with respect to the Riemannian metric. Recall that in the proximal view of GD, we have:

Algorithm 2: Proximal Point method

Given a function $f\colon \mathbb{R}^d \to \mathbb{R}$ and step size $\eta > 0$, compute:

$$x_{t+1} = \operatorname{prox}_{h_t,\eta}(x_t), \quad t = 1, 2, \ldots$$

where $h_t\colon \mathbb{R}^d \to \mathbb{R}$ are convex functions such that

$$\forall x \in \mathbb{R}^d\colon h_t(x) \le f(x) \le h_t(x) + \frac{\beta}{2}\|x - x_t\|^2.$$

That is:

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x_t) + \nabla f(x_t) \cdot (x - x_t) + \frac{1}{2\eta}\|x - x_t\|^2 \right\}$$

and in MD instead of η as a denominator of the distance, we control the ratio by a coefficient η of the gradient.

Theorem 2.2.1. $x(t)$ ***exists and is unique***
The solution to the Mirror Descent algorithm exists and is unique.

Proof. $x(t)$ **exists and is unique**

$$\begin{aligned}
D_\phi(y, x(t)) &= \phi(y) - \phi(x(t)) - \nabla\phi(x(t)) \cdot (y - x(t)) \\
\frac{d}{dt} D_\phi(y, x(t)) &= -\nabla\phi(x(t)) \cdot \frac{d}{dt}x(t) - \nabla^2\phi(x(t))\frac{d}{dt}x(t) \cdot (y - x(t)) - \nabla\phi(x(t)) \cdot (-\frac{d}{dt}x(t)) \\
&= -\nabla^2\phi(x(t)) \cdot \frac{d}{dt}x(t) \cdot (y - x(t)) \quad (-[\nabla^2\phi(x(t))] \cdot \frac{d}{dt}x(t) = (\eta g(t) + \lambda(t))) \\
&= (\eta g(t) + \lambda(t)) \cdot (y - x(t)) \quad (y \in \mathcal{K} \text{ and } \lambda(t) \in N_{\mathcal{K}}(x(t)) \Rightarrow \lambda(t) \cdot (y - x(t)) \le 0) \\
&\le \eta g(t) \cdot (y - x(t))
\end{aligned}$$

□

We can think that $g(t) \cdot y$ is what OPT would have pay at time t, and $g(t) \cdot x(t)$ is what we pay. If we pay more than OPT, then the Bregman divergence will decrease and we will move towards OPT. We get that for any $y \in \mathcal{K}$:

$$\int_0^T g(t) \cdot (x(t) - y)dt \le \frac{D_\phi(y, x(0))}{\eta}$$

- Previously we had 2 terms in the RHS - one for the range and one for the variance, now we have only one term for the range, because we are in a continuous time setting).
- By itself, this theorem is meaningless, we can always speed up the time by making ϕ smaller (and the Bregman divergence smaller).

Discrete time version:

$$\begin{aligned}
&\text{discretize } \frac{d}{dt}x(t) = -[\nabla^2\phi(x(t))]^{-1} \cdot g(t) \iff \nabla^2\phi(x(t))\frac{d}{dt}x(t) = -g(t) \\
&\iff \frac{d}{dt}\nabla\phi(x(t)) = -g(t)
\end{aligned}$$

Then we get that:

$$\nabla\phi(x(t+1)) = \nabla\phi(x(t)) - \eta g(t)$$

In GD, we have that $x_{t+1} = x_t - \eta\nabla f(x_t)$, and in MD we have that $\nabla\phi(x_{t+1}) = \nabla\phi(x_t) - \eta g(t)$. We can think of it as - instead of doing an update in the primal space, we do an update in the dual space.

Moreover, GD didn't really make sense since x_t was a primal point and $\nabla f(x_t)$ was a dual point (it was a linear functional) - so we added primal and dual points, which was ok just in finite dimensions.

Mirror Descent fixes this issue - it first map the primal point to the dual space, then do the update in the dual space, and then map back to the primal space.

Chapter 3

The Third Miracle:

3.1 Coupon Collector Phenomenon

The coupon collector's problem is a classic example of a problem in probability theory that has applications in various fields such as computer science, combinatorics, and stochastic processes. The problem describes a scenario where a collector wishes to gather one of each type of coupon, and each coupon type is equally likely to be obtained at each trial.

Definition 3.1.1. ***Coupon Collector Problem***
The Coupon Collector Problem *considers a scenario with n different types of coupons.*
Each time a coupon is collected, it is of a type chosen uniformly at random from the n types.
The problem is to determine the expected number of coupons that must be collected before having at least one of each type.

Theorem 3.1.1. ***Expected Number of Trials for Coupon Collection***
Let T be the random variable representing the number of trials needed to collect all n types of coupons.
Then the expected value of T is $E[T] = n \cdot H_n$, where H_n is the n-th harmonic number defined as $H_n = \sum_{k=1}^{n} \frac{1}{k}$.

Proof. The collection process can be broken into n stages, where stage i represents the collection of the i-th new coupon after already having $i-1$ different types. The expected number of trials to collect a new coupon during stage i is $\frac{n}{n-(i-1)}$, since there are $n-(i-1)$ types not yet collected (geometric distribution).
Therefore, the expected total number of trials to collect all n coupons is:

$$E[T] = \sum_{i=1}^{n} \frac{n}{n-(i-1)} = n \cdot \sum_{i=1}^{n} \frac{1}{i} = n \cdot H_n.$$

Hence, the expected number of trials needed to collect all n types of coupons is n times the n-th harmonic number (and is O(n log n)). □

3.2 Metrical Task Systems (MTS)

Definition 3.2.1. ***MTS (Metrical Task System)*** *[Borodin, Linial, Saks, '92]*
Let X be a finite metric space.
The player maintains a state $\rho_t \in X$. At time $t+1$ the player receives a task $c_{t+1} : X \to \mathbb{R}_+$ (a cost function) and move to a new state $\rho_{t+1} \in X$.
The movement cost = $dist(\rho_t, \rho_{t+1})$.
The service cost = $c_{t+1}(\rho_{t+1})$.

Some applications:

- **Self adjusting data structures** - the state is the data structure, the task is the query, the cost is the time to answer the query.
  For example: $X = S^n$ set of permutations of n elements,
  cost function: $c_{t+1}(\sigma) = \sum_{i=1}^{n'} \mathbb{1}(\sigma(i) \neq j)$ and $\sigma(n') = j$
- **Power management** - the states are On / Off / Sleep.
  $dist(\text{On}, \text{Off}) > dist(\text{On}, \text{Sleep})$.
  $c_1(\text{On}) = 0, c_1(\text{Sleep}) = c_1(\text{Off}) = \infty$.
  $c_2(\text{On}) = 1, c_2(\text{Sleep}) = \epsilon, c_2(\text{Off}) = 0$.
- Strictly more general than **prediction with expert advice** -
  Imagine that the metric space is trivial (i.e. $dist(x, y) = 0$ for all x, y such that $x \neq y$).
  If I don't move, than being evaluated at the next cost function or the current function is the same. The cost that I get when I'm evaluated at the next cost function is upper bounded by the cost that I'm evaluated at the current cost function + the distance that I moved.

The type of guarantees:

- **Competitive ratio** $\equiv \frac{ALG}{OPT}$ - the ratio between the cost of the algorithm and the cost of the best fixed states in hindsight.
- **Conjecture in online algorithms** - for any metric space $|X| = n, \quad \exists \quad \mathcal{A}$ in $O(\log n)$-competitive algorithm. Where the $\log n$ comes from? We have seen that in the experts problem, the optimal rate is $\sqrt{T \log n}$ (also exists in matrix concentration results).

Today's Theorem [Bubeck, Cohen, J.Lee, YT.Lee, '18] - $O(\log^2 n)$-competitive algorithm for MTS. Remember that in some example it is $O(\log n)$ where n is really the parameter of interest, but in other examples n is exponential is the parameter of interest (and then we get $O(\log^2 n)$). $\forall X$ metric space, $\Omega(\frac{\log n}{\log \log n})$ is necessary.

3.2.1 Deterministic Strategies on uniform metric

By "uniform metric" we mean that the distance between any two points is 1.
For any deterministic algorithm, it is at least in order of n since the adversary can choose the highest cost function for the algorithm.
Algorithm in $O(n)$-competitive:
Stay in state until it accumulates a cost of 1. Then mark the state and move to an unmarked state. Once everything is marked, I paid $\leq$ 2n and OPT pays at least 1 (since every state is evaluated at least once). So I'm 2n-competitive.
Why is it $\Omega(n)$-competitive?
$c_{t+1}(\rho) = +\infty \mathbb{1}_{\rho = \rho_t}$. Over n time steps, the algorithm pays n and OPT pays less than 1, so the competitive ratio is at least n.

3.2.2 Randomized Strategies for Uniform Metric

Randomized algorithms, essential in adversarial models, create uncertainty to prevent adversaries from precisely exploiting the algorithm's state. The adversary, lacking knowledge of the algorithm's random location, is forced to target randomly, making the cost sequence unpredictable.
Such randomness leads to bounded accumulated costs over n iterations. The strategy in a uniform metric space dictates that an algorithm stays in a state until accruing a specified cost,

then moves to a random unmarked state. The expected movement function $f(n)$, where n states are unmarked, is given by:

$$\begin{aligned} f(n) &= \text{the expected number of moves when } n \text{ states are unmarked}, \\ &= \frac{1}{n} + f(n-1), \\ &= 1 + \frac{1}{2} + \ldots + \frac{1}{n}, \\ &= \log n + O(1), \\ &= O(\log n). \end{aligned}$$

Imagine we have n states left to mark, what's the probability that our current state will be the first one to accumulate a cost of 1? $\frac{1}{n}$. If we are not the first, then we have $n-1$ states left to mark, and the probability that we will be the first to accumulate a cost of 1 is $\frac{1}{n-1}$. Thus, the expected movements decrease logarithmically with the number of unmarked states.
What about the lower bound?
$c_{t+1}(\rho) = +\infty$ at a random location. ALG's cost over n time is 1 in expectation (expectation of geometric distribution with parameter $\frac{1}{n}$), but it takes $O(nlogn)$ to cover X (to hit all the states (the coupon collector problem)). So the competitive ratio is $\Omega(\log n)$.

Open question: "weighted coupon collector problem" - the cost of collecting a coupon is not 1, but a given weight (the cost function is not uniform).

3.3 Beyond Uniform Spaces: weighted stars

Continues time problem: instead of $c_t, \quad t \in \mathbb{N}$, we have $c_t, \quad t \in \mathbb{R}_+$.

Lemma 3.3.1. *Continues time guarantees transfer to discrete time guarantees*

Proof. $c_t \to c(t)$ is a water filling process.
from $\rho_t \to \rho(t)$, which is the smallest cost state visited by the continuous time algorithm.
SC-DC ≤ SC-CS
MC-DC ≤ MC-CS
in general OPT-CT ≤ OPT-DC but by the two inequalities above, we get that OPT-CT ≤ OPT-CS. ☐

3.3.1 Wasserstein Distance

The Wasserstein distance is a measure of the distance between two probability distributions that has gained popularity in various fields, particularly in machine learning for its application in generative models.

Definition 3.3.1. ***Wasserstein Distance***
Given two probability measures μ and ν on a metric space (M, d), the Wasserstein distance of order p between μ and ν is defined as:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{M \times M} d(x, y)^p \, d\gamma(x, y) \right)^{\frac{1}{p}},$$

where $\Gamma(\mu, \nu)$ denotes the collection of all measures on $M \times M$ with marginals μ and ν on the first and second factors, respectively.

Intuition: The Wasserstein distance can be thought of as the minimum "cost" required to transform one distribution into another, where the cost is measured by the amount of "mass" that must be moved times the distance it has to be moved.
Example: Consider two one-dimensional distributions, μ which is a delta function at 0, and ν which is a delta function at 1. The Wasserstein distance of order 1 between μ and ν is simply the distance between the two points, which is 1.
This metric is particularly useful when comparing probability distributions that might not have the same support or when the cost of transportation is significant.

3.3.2 Continues space problem

Instead of $\rho_t \in X$, we maintain $x_t \in \triangle_n$ (simplex).
$dist(\rho_t, \rho_{t+1}) \rightsquigarrow w_1(x_t, x_{t+1}) := \inf_{(X,Y):X\sim x_t, Y\sim x_{t+1}} \mathbb{E}dist(X, Y)$.

Chapter 4

The Fourth Miracle:

Chapter 5

The Fifth Miracle: