$O(1/\epsilon)$ Lower Bounds for Convex Optimization with Stochastic Oracles

Yaoliang Yu YAOLIANG@CS.UALBERTA.CA
Xinhua Zhang XINHUA.ZHANG.CS@GMAIL.COM

Department of Computing Science
University of Alberta T6G 2E8, Canada

This note shows that with a stochastic oracle, there exists a strongly convex objective such that no deterministic optimization algorithm can reduce the expected gap at a rate faster than $O(1/\epsilon)$.

We first formalize stochastic optimization in the oracle-versus-optimizer paradigm (Nemirovski and Yudin, 1983) in Section 1, and then sketch the state-of-the-art upper and lower bounds for the rate of convergence (Agarwal et al., 2009) in Section 2. Intuitively, they show that there exists a first-order stochastic oracle (which returns a noisy version of the gradient with zero mean and bounded variance), such that for any optimization algorithm there exists an objective function on which the gap is at least $\Omega(1/t)$ after querying the oracle for t times. However, this result is weak in the following sense:

- They contrived a tough oracle which makes optimization hard. But in practice the optimizer *can* engineer the stochastic mechanism in the oracle.
- The oracle was restricted to first-order, *i.e.* the expectation of the randomly returned result is the true gradient. It is well known that by using a more expensive/informative oracle such as the expectation oracle introduced by Nesterov (2005b), the rate of convergence for deterministic algorithms can be accelerated from $O(1/t)$ to $O(1/t^2)$. Therefore, it is conceivable that using a stochastic version of such an oracle can also yield $O(1/t^2)$ rate of convergence.
- The outcome of the oracle is used for explicit updates. However, in practice implicit update is often observed to converge faster.

Therefore, it was our hope that the $O(1/t)$ lower bound could be bypassed by using implicit update, and/or more informative oracles with a better stochastic mechanism. Unfortunately, this hope was later dismissed in Section 3 by using a simple one dimensional example. By making a weak assumption on the optimizer which is often satisfied, we strengthened the minimax lower bound in (Agarwal et al., 2009) into a maximin (uniform) lower bound.

1. Oracle-versus-optimizer Model for Stochastic Optimization

In this section we try to give a rigorous and self-contained formalization of stochastic optimization. Although the ideas are sometimes quite obvious, many subtleties exist and various normalizations are needed to rule out vacuous results. Nemirovski and Yudin (1983) devoted one chapter to this end, and we just highlight some important conceptualizations based on (Agarwal et al., 2009).

Suppose we have a class of (deterministic) convex functions $\mathcal{F} = \left\{f : \mathbb{R}^d \mapsto \mathbb{R} : f \text{ is convex}\right\}$, and a class of feasible regions $\mathbb{S} = \left\{S \subseteq \mathbb{R}^d : S \text{ is convex}\right\}$. Given any function $f \in \mathcal{F}$ and feasible region $S \in \mathbb{S}$, we want to find its minimizer

$$\min_{\mathbf{w} \in S} f(\mathbf{w}).$$

An optimization method is any procedure that solves this task, typically by repeatedly selecting values from S. Our primary focus is: given some $\mathcal{F}$ and $\mathbb{S}$, what is the minimum computational labor any such optimization method would expend to find an ϵ accurate solution for any $f \in \mathcal{F}$ and $S \in \mathbb{S}$? An ϵ accurate solution $\bar{\mathbf{w}}$ means $f(\bar{\mathbf{w}}) \le \min_{\mathbf{w} \in S} f(\mathbf{w}) + \epsilon$. We call $f(\bar{\mathbf{w}}) - \min_{\mathbf{w} \in S} f(\mathbf{w})$ the *gap* of $\bar{\mathbf{w}}$.

In order to address this question, we follow the oracle model abstracted by Nemirovski and Yudin (1983) which encompasses the following key elements.

- A class of objective functions $\mathcal{F}$ and feasible regions $\mathbb{S}$,
- A set of deterministic or stochastic oracles $\mathcal{O}$,
- For a given integer t, a class $\mathbb{M}_t$ of methods making t oracle queries.

Objective function and feasible region Scaling has subtle effect on the lower bounds. Suppose by applying a method $\mathcal{M}$ to an objective function f constrained to feasible region S, the gap can be reduced to $\frac{|f(\mathbf{w}_0)|}{k}$ after making k oracle queries where $\mathbf{w}_0$ is the initial guess. Then by arbitrarily scaling up f, the gap can be arbitrarily scaled up. Similarly, when f is unbounded on S, changing $\mathbf{w}_0$ also allows one to arbitrarily scale up $f(\mathbf{w}_0)$. Therefore, scaling assumptions are often introduced such that, among other consequences, $\max_{\mathbf{w} \in S} |f(\mathbf{w})| < L < \infty$ for all $f \in \mathcal{F}$ and $S \in \mathbb{S}$. For example, Agarwal et al. (2009) assumed that

$$\mathcal{F} = \left\{f : \mathbb{R}^d \mapsto \mathbb{R} : f \text{ is 1-Lipschitz continuous, } i.e. \; |f(\mathbf{x}) - f(\mathbf{y})| \le \|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{x}, \mathbf{y} \in S\right\} \quad (1)$$

$$\mathbb{S} = \left\{S \subseteq \mathbb{R}^d : S \text{ is convex, } \|\mathbf{x} - \mathbf{y}\|_\infty \le 1 \text{ for all } \mathbf{x}, \mathbf{y} \in S\right\}. \quad (2)$$

One can also assume that the nonsmooth objective f contains a bounded subgradient everywhere:

$$\mathcal{F} = \left\{f : \forall\ \mathbf{w}, \exists\ \mathbf{g} \in \partial f(\mathbf{w}), \text{ s.t. } \|\mathbf{g}\| \le 1\right\}.$$

A weaker assumption made by Abernethy et al. (2008) is that the f contains a bounded subgradient at every iterate $\mathbf{w}_t$. Since $\mathbf{w}_t$ depends on the optimization method, this assumption in fact couples the optimization method and the problem. Assumptions can also be be made indirectly. For example, a standard assumption made in machine learning is that the data points $\mathbf{x}_i$ are contained in a R radius ball. This indirectly introduces restrictions on the empirical risk $\ell(\mathbf{w}; \mathbf{x}_i, y_i)$.

One can also restrict the choice of $\mathbb{S}$. In fact, the dependence of the lower bound on the shape of the feasible region S can be quite interesting (via the so-called asphericity).

Oracle An oracle is a possibly random function ϕ which takes a function f and a query point $\mathbf{w}$, and responses with an element $\phi(\mathbf{w}, f)$ in an information set $\mathcal{I}$. For example, an exact oracle of k-th order returns $f(\mathbf{w})$ and the first k derivatives of f at $\mathbf{w}$. We now only consider first order oracle and if f is not differentiable at $\mathbf{w}$ then any arbitrary subgradient of f at $\mathbf{w}$ is returned in place. For the case of stochastic oracles which is the focus of this note, these values can be corrupted with zero-mean noise whose variance is bounded. Later, we will discuss in more depth the stochasity and localness of the oracle.

Method Method models the iterative optimization algorithm, which maps the outcome of the oracle from step 1 to $t-1$ to a new iterate $\mathbf{w}_t$. If the oracle is stochastic, which means its result is random, then $\mathbf{w}_t$ is also random.

1.1 Definition of rate of convergence

Although we have been discussing rates for lunch, its meaning can be quite tricky and a rigorous formalization is important. Otherwise it is very easy to draw plausible conclusions.

In our oracle/method model, the rate of convergence coincides with the oracle complexity, *i.e.* the number of queries made to the oracle. We will stick to this definition in the whole note.

Given some number of rounds/iterations t, an optimization method $\mathcal{M}_t$ designed to approximately minimize a function f over a convex set S proceeds as follows: at any round $k = 1, ..., t$, the method $\mathcal{M}_t$ queries the oracle ϕ at $\mathbf{w}_k \in S$, and the oracle returns the information $\phi(\mathbf{w}_k, f)$. The method then uses the information $\phi(\mathbf{w}_1, f), \ldots \phi(\mathbf{w}_k, f)$ to decide at which point $\mathbf{w}_{k+1}$ the next query should be made. We only consider deterministic methods, *i.e.* this mapping to $\mathbf{w}_{k+1}$ is not random. For a given oracle ϕ, let $\mathbb{M}_t$ denote the class of all optimization methods $\mathcal{M}_t$ that make t queries according to the procedure outlined above. For any method $\mathcal{M}_t \in \mathbb{M}_t$, we denote its error on function f after t steps as

$$\epsilon(\mathcal{M}_t, f, S, \phi) := f(\mathbf{w}_{t+1}) - \inf_{\mathbf{w}\in S} f(\mathbf{w}). \tag{3}$$

When the oracle is stochastic, $\mathbf{w}_{t+1}$ is random and so is $\epsilon(\mathcal{M}_t, f, S, \phi)$. Then a natural measurement of accuracy is in terms of the expected value

$$\mathbb{E}_\phi \left[\epsilon(\mathcal{M}_t, f, S, \phi)\right]. \tag{4}$$

where the expectation is taken over the randomness of the *outcome* of the oracle.

We can consider that $\mathbb{M}_t$ is trying to accelerate the optimization by choosing the best method in it, while $\mathcal{F}$, $\mathbb{S}$ and $\mathcal{O}$ are trying to find a tough combination of their elements to slow down the optimization. In practice, a tuple of elements from $\mathcal{F}$ and $\mathbb{S}$ defines an optimization problem. For a fixed oracle $\phi \in \mathcal{O}$ we can define a minimax gap and a maximin gap:

$$\begin{aligned} \text{(minimax gap)} \quad \epsilon^*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi) &:= \inf_{\mathcal{M}_t\in\mathbb{M}_t} \sup_{S\in\mathbb{S}} \sup_{f\in\mathcal{F}} \mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)], && (5) \\ \text{(maximin gap)} \quad \epsilon_*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi) &:= \sup_{S\in\mathbb{S}} \sup_{f\in\mathcal{F}} \inf_{\mathcal{M}_t\in\mathbb{M}_t} \mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)]. && (6) \end{aligned}$$

They are both functions of t and ϕ, and of other hyperparameters used to define $\mathcal{F}$, $\mathbb{S}$, and $\mathbb{M}_t$. The name "minimax" refers to $\inf_{\mathcal{M}_t} \sup_{f,S}$, and "maximin" refers to $\sup_{f,S} \inf_{\mathcal{M}_t}$. It is well known that the minimax gap majorizes maximin gap:

$$\epsilon^*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi) \geq \epsilon_*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi). \tag{7}$$

There are four possible bounds: upper/lower bounding the minimax/maximin gap. We start from minimax gap.

Minimax upper bound The minimax upper bound simply upper bounds the minimax error defined by (5):

$$\epsilon^*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi) = \inf_{\mathcal{M}_t \in \mathbb{M}_t} \sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)] \leq u^*(t; \phi), \quad \forall\, t \in \mathbb{N}.$$

It means fixing t and ϕ, there exists a method $\mathcal{M}_t^*$ such that for any problem (f, S), $\mathcal{M}_t^*$ can reduce the expected error to be below $u^*(t; \phi)$ in t steps.

$$\boxed{\text{Fixing } \phi,\ \forall\, t,\ \exists\, \mathcal{M}_t^*,\ s.t.\ \forall\, (f, S),\ \mathbb{E}_\phi[\epsilon(\mathcal{M}_t^*, f, S, \phi)] \leq u^*(t; \phi).}$$

Here $u^*(t; \phi)$ is a function of t parameterized by ϕ, $\mathbb{M}_t$, $\mathcal{F}$ and $\mathbb{S}$. It does not depend on the particular problem (f, S), but uniformly for the class $\mathcal{F}$ and $\mathbb{S}$. This will require restricting the scale of $\mathcal{F}$ and $\mathbb{S}$ as was discussed above.

Minimax lower bound One can also lower bound the minimax gap:

$$\epsilon^*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi) = \inf_{\mathcal{M}_t \in \mathbb{M}_t} \sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)] \geq l^*(t; \phi), \quad \forall\, t \in \mathbb{N}$$

where $l^*(t; \phi)$ is a positive valued function of t parameterized by ϕ and the hyperparameters used to define $\mathcal{F}$, $\mathbb{S}$, and $\mathbb{M}_t$. This minimax lower bound says intuitively that fixing the oracle, for any fixed number of query t, no matter what method (optimizer) $\mathcal{M}_t \in \mathbb{M}_t$ is used, there must exist a problem $(f, S) \in \mathcal{F} \times \mathbb{S}$ such that the expected gap/error is at least $l^*(t; \phi)$.

$$\boxed{\text{Fixing } \phi,\ \forall\, t,\ \forall\, \mathcal{M}_t^*,\ \exists\, (f, S),\ s.t.\ \mathbb{E}_\phi[\epsilon(\mathcal{M}_t^*, f, S, \phi)] \geq l^*(t; \phi).}$$

In this sense, the game between the problem and solver runs as follows:

- Player (solver) picks an optimization algorithm $\mathcal{M}_t \in \mathbb{M}_t$,
- Adversary picks a tough pair of objective function $f \in \mathcal{F}$ and feasible region $S \in \mathbb{S}$,
- Then the player (solver) suffers at least some amount of expected error.

Since the solvers plays first and the adversary can customize a tough problem against this particular optimization algorithm, so we can expect that the solver will suffer a high expected error.

Maximin (uniform) lower bound We can also lower bound the maximin gap as follows:

$$\epsilon_*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi) := \sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \inf_{\mathcal{M}_t \in \mathbb{M}_t} \mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)] \geq l_*(t; \phi), \quad \forall\, t \in \mathbb{N}. \tag{8}$$

This maximin lower bound says intuitively that fixing the oracle, for any number of query t, there exists a problem $(f, S) \in \mathcal{F} \times \mathbb{S}$ such that all methods will suffer the expected gap at least $l_*(t; \phi)$.

$$\boxed{\text{Fixing } \phi,\ \forall\, t,\ \exists\, (f, S),\ s.t.\ \forall\, \mathcal{M}_t^*,\ \mathbb{E}_\phi[\epsilon(\mathcal{M}_t^*, f, S, \phi)] \geq l_*(t; \phi).}$$

In this sense, the game between the problem and solver runs as follows:

- Adversary picks a tough objective function $f \in \mathcal{F}$ and feasible region $S \in \mathbb{S}$,
- Player (solver) picks an optimization algorithm $\mathcal{M}_t \in \mathbb{M}_t$,
- Then the player (solver) suffers at least some amount of expected error.

Since the adversary plays first and the solver can choose an algorithm that is most efficient for this particular problem, we may expect that the solver will suffer a lower expected gap. We also call maximin lower bound as "uniform lower bound" because it finds a tough problem which is hard for all solvers $\mathcal{M}_t$.

Maximin upper bound Simply upper bound the maximin gap by

$$\epsilon_*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi) := \sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \inf_{\mathcal{M}_t \in \mathbb{M}_t} \mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)] \leq u_*(t; \phi), \quad \forall\, t \in \mathbb{N}. \tag{9}$$

This maximin upper bound says that fixing the oracle, for any number of query t and any problem $(f, S) \in \mathcal{F} \times \mathbb{S}$, there exists a method which reduces the expected gap to at most $u_*(t; \phi)$.

$$\boxed{\text{Fixing } \phi,\ \forall\, t,\ \forall\, (f, S),\ \exists\, \mathcal{M}_t^*,\ s.t.\ \mathbb{E}_\phi[\epsilon(\mathcal{M}_t^*, f, S, \phi)] \leq u_*(t; \phi).}$$

Clearly this is weaker than the minimax upper bound which says there exists a method which can minimize all problems efficiently.

$1/\epsilon$ and $1/k$ representation It is well known that the $1/t$ rate is equivalent to the $1/\epsilon$ rate, and the above t based definition can also be translated into the language of ϵ. Which style of definition to use depends on the convenience of analysis, usually affected by the class of algorithms $\mathcal{M}$. For example, Nesterov (2005a) requires a prescribed gap ϵ, *i.e.* only an ϵ accurate solution is required. Then it uses ϵ to construct a smooth convex function that approximates the original objective function, such that their difference lies uniformly in $[0, \epsilon]$. Afterwards any solver can be applied to optimize this surrogate objective. On the other hand, some real applications have a notion of budget, *i.e.* the maximum number of steps that the solver is allowed to take (or the number of times that the oracle can be invoked). For example, some solvers can reduce the gap to a medium level (*e.g.* 10^{-3}) in just a handful of steps, but is very inefficient in finding a high precision solution. In contrast, some solvers has a good local rate of convergence (in the vicinity of the solution) but needs a lot of steps to find such a vicinity. So in the budget based setting, it appears more convenient to use t based analysis.

Lower and upper bounds for a *fixed* method Sometimes we have already engineered a method $\mathcal{M}_t$, and it is our goal to analyze its efficiency. In this case, we are interested in

$$\epsilon(\mathcal{M}_t, \mathcal{F}, \mathbb{S}, \phi) := \sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)]. \tag{10}$$

However there is still a minimax hidden in the definition if we recall that the step index t is fixed prior to the two sup operations in (10). In other words, the argsup can change with t. So

$$\sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)] \geq l^*(t; \phi) \quad \text{for all } t \tag{11}$$

means that for any t, there exists a problem (f, S) which depends on t such that the expected gap is at least $l^*(t; \phi)$. Here $l^*(t; \phi)$ is a function of t parameterized by ϕ, $\mathcal{F}$, and $\mathbb{S}$. (11) can be easily rewritten as

$$\inf_t \sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \frac{\mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)]}{l^*(t; \phi)} \geq 1.$$

which makes the minimax explicit. To derive a uniform variant, we just need to change the minimax into maximin:

$$\sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \inf_t \frac{\mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)]}{l_*(t; \phi)} \geq 1,$$

which means there exists a problem (f, S) such that for all t, $\mathcal{M}_t$ suffers an expected gap at least $l_*(t; \phi)$. As for upper bound, we simply use

$$\sup_t \sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \frac{\mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)]}{u(t; \phi)} \leq 1 \qquad \textit{i.e.}, \quad \epsilon(\mathcal{M}_t, \mathcal{F}, \mathbb{S}, \phi) \leq u(t; \phi) \text{ for all } t. \tag{12}$$

That is, for any problem (f, S) and any t, the expected gap incurred by $\mathcal{M}_t$ is at most $u(t; \phi)$.

1.2 Relationship between different bounds and tightness

Now that the maximin gap $\epsilon_*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi)$ minorizes the minimax gap (Eq (7)), the maximin lower bound is stronger than the minimax lower bound because any lower bound of $\epsilon_*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi)$ must also be a lower bound of $\epsilon^*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi)$ (the converse is not necessarily true). Similarly, the minimax upper bound is stronger than the maximin upper bound because any upper bound of $\epsilon^*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi)$ must also be an upper bound of $\epsilon_*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi)$.

Since both minimax upper bound and minimax lower bound use the minimax gap $\epsilon^*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi)$, it is clear that if one can upper and lower bound $\epsilon^*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi)$ by $O(c(t))$ and $\Omega(c(t))$ respectively, then the analysis must be tight on both sides. This means to study the tightness of the minimax upper bound, there is no need to examine the maximin lower bound which is usually much harder to establish than the minimax lower bound. But of course, maximin lower bound is interesting and useful in its own right, and is useful for studying the optimality of maximin upper bound.

When the optimization method $\mathcal{M}_t$ is fixed, it is also clear that if one can upper and lower bound $\epsilon^*(\mathcal{M}_t, \mathcal{F}, \mathbb{S}, \phi)$ by $O(c(t))$ (in (12)) and $\Omega(c(t))$ (in (11)) respectively, then the analysis must be tight on both sides.

1.3 A variety of ways to qualify the oracle

In the above we deliberately left the oracle ϕ unqualified. But it is sometimes meaningful to qualify it universally on top of the minimax gap, which is the approach taken by Agarwal et al. (2009):

$$\epsilon^*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \mathcal{O}) := \sup_{\phi \in \mathcal{O}} \inf_{\mathcal{M}_t \in \mathbb{M}_t} \sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \mathbb{E}_{\phi}[\epsilon(\mathcal{M}_t, f, S, \phi)]. \tag{13}$$

In this sense, the game between the problem and solver proceeds as follows after fixing t:

- Adversary picks a tough oracle $\phi \in \mathcal{O}$,
- Player (solver) picks an optimization algorithm $\mathcal{M}_t \in \mathbb{M}_t$,
- Adversary picks a tough pair of objective function $f \in \mathcal{F}$ and feasible region $S \in \mathbb{S}$,
- Finally the player (solver) suffers some amount of expected error.

However, taking sup over the oracle is not realistic in stochastic convex optimization, because in reality the oracle is not necessarily picked against the solver. In many applications, the solver has the right to specify the oracle. This motivates the use of taking $\inf_{\phi \in \mathcal{O}}$ for both $\epsilon^*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi)$ and $\epsilon_*(\mathbb{M}_t, \mathcal{F}, \mathbb{S}, \phi)$. However, this will be too optimistic because the solver can hardly tell which oracle is the best either. So in practice, it will just choose some default distribution, which typically lies between the worst and the best oracle. That is why we leave the oracles unqualified in the definition (5) and (9), and it is meaningful to study the lower bound for different choices of ϕ.

1.4 More details on oracle: deterministic and stochastic

In this section, we discuss in more depth the concept of oracle. A good understanding is critical for us to bypass the existing lower bounds given by, *e.g.*, (Agarwal et al., 2009) and (Abernethy et al., 2008).

1.4.1 Deterministic oracles

In deterministic class of oracles, although the k-th order oracle is probably the most commonly used, some other oracles are also possible and could yield faster convergence. For example, consider the objective function

$$f(\mathbf{w}) = \max_i \langle \mathbf{a}_i, \mathbf{w} \rangle + b_i.$$

Given a query point $\mathbf{w}_t$, the first order deterministic oracle will return the subgradient

$$\mathbf{g}_t := \mathbf{a}_{i^*}, \quad \text{where } i^* = \operatorname*{argmax}_i \langle \mathbf{a}_i, \mathbf{w}_t \rangle + b_i \tag{14}$$

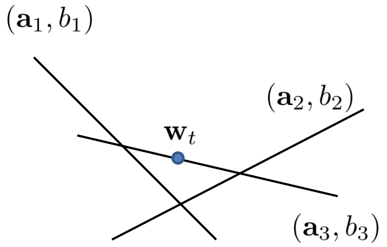

Figure 1: Illustration of different oracles on a piecewise linear function. At $\mathbf{w}_t$, the first order oracle returns $\mathbf{a}_3$. The exp oracle returns a weighted average of $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$.

In other words, the returned subgradient does not change even if we perturb $\mathbf{a}_j$ and b_j for all $j \neq i^*$ (so far as i^* is still the argmax). Another more soft oracle will return

$$\tilde{\mathbf{g}}_t := \sum_i \mathbf{a}_i \frac{\exp\left\{\frac{1}{T}\left(\langle \mathbf{a}_i, \mathbf{w}_t \rangle + b_i\right)\right\}}{\sum_j \exp\left\{\frac{1}{T}\left(\langle \mathbf{a}_j, \mathbf{w}_t \rangle + b_j\right)\right\}}, \quad \text{where } T > 0. \tag{15}$$

So $\tilde{\mathbf{g}}_t$ is a weighted average of all $\mathbf{a}_i$, and $\mathbf{a}_i$ receives a higher weight if $\langle \mathbf{a}_i, \mathbf{w}_t \rangle + b_i$ is large. Hence we call it "expectation oracle" or simple "exp oracle". T serves as a temperature which controls the softness. If the argmax in (14) is unique, then $\tilde{\mathbf{g}}_t$ tends to $\mathbf{g}_t$ when the temperature T approaches 0. See Figure 1 for an illustration. It is noteworthy that the lower bounds under k-th order oracle can often be readily extended to the so-called "local oracle", which roughly speaking, depends only on the function *value* (rather than its expression) in some small neighborhood of $\mathbf{w}_t$. Clearly, the exp oracle is not local, because adding or removing the line $(\mathbf{a}_2, b_2)$ in Figure 1 does not change the function value at the neighborhood of $\mathbf{w}_t$, while it does change the result of the exp oracle.

1.4.2 Stochastic oracles

The motivation of using stochastic oracles is basically two folds: computational efficiency and oracle viability. Although we have been bounding the number of oracle queries, the cost for the oracle to answer each query definitely matters in practice. The main advantage of stochastic optimization is that the oracle is often computationally cheap. For example, consider the objective function (in $\mathbf{w}$):

$$\mathop{\mathbb{E}}_{\alpha \sim Q}[f(\mathbf{w}; \alpha)], \qquad \text{where} \quad \alpha \text{ has a distribution } Q. \tag{16}$$

Assume that all regularities conditions hold (*e.g.*, swapping expectation and gradient). A stochastic oracle may respond to query $\mathbf{w}_t$ by

$$f(\mathbf{w}_t; \alpha) \text{ and } \nabla_{\mathbf{w}} f(\mathbf{w}_t; \alpha) \qquad \text{where} \quad \alpha \text{ is drawn from } Q.$$

Another situation of using stochastic optimization is when the expectation in (16) is computationally intractable, or even when Q is unknown (but samples can be drawn).

Although the outcome of the oracle is not the exact function value or gradient, many algorithms can use this noisy information to asymptotically find a high precision solution with high probability.

Obviously, using a stochastic oracle will require more queries than deterministic oracles to achieve a fixed level of accuracy. But this is compensated by its much lower cost per invocation. More importantly, stochastic oracle often still allows one to efficiently find a low precision solution. This is usually enough for machine learning purposes because the objective function is just an approximate model in the first place (see (Bottou and Bousquet, 2007) for the tradeoff between approximation error, estimation error and optimization error).

Despite the randomness of the stochastic oracle, conceivably it is should still satisfy some statistical conditions. Formally, let there be a random variable ω which controls the stochastic oracle, *i.e.* once the observation of ω is available, the result of the oracle can be determined with no more randomness. Let the objective function be $f(\mathbf{w})$. We follow Nemirovski and Yudin (1983) and denote the oracle result as $(f(\mathbf{w};\omega), f'(\mathbf{w};\omega))$ which slightly abuses the notation. Now we can formally express the "standard" conditions of stochastic oracles: unbiasedness and bounded variance:

$$\begin{aligned} \text{Unbiasedness} \quad &: \quad \mathbb{E}_\omega[f(\mathbf{w};\omega)] = f(\mathbf{w}), \text{ and } \mathbb{E}_\omega[f'(\mathbf{w};\omega)] = \nabla f(\mathbf{w}); &(17)\\ \text{Bounded variance} \quad &: \quad \exists L > 0,\ s.t.\ \mathrm{Var}_\omega[f'(\mathbf{w};\omega)] \preceq LI \text{ and } \mathrm{Var}_\omega[f(\mathbf{w};\omega)] < L. &(18)\end{aligned}$$

Here I is the identity matrix and $A \preceq B$ means the matrix $B - A$ is positive semi-definite. These conditions on "oracle" are effectively conditions on the distribution over ω and the mechanism that ω deterministically assigns the result of oracle. Using this notation, the iterate $\mathbf{w}_{t+1}$ is a deterministic function of the latent variable $\omega_1^t := (\omega_1, \ldots, \omega_t)$ through the last t steps, and the expected gap at step $t+1$ can be expressed as the expectation over ω_1^t: $\epsilon_{t+1} := \mathbb{E}_{\omega_1^t}[f(\mathbf{w}_{t+1}) - \min_{\mathbf{w}} f(\mathbf{w})]$. In general, $\omega_1, \ldots, \omega_t$ can be dependent, but we assume they are *iid* throughout the whole note.

The stochastic oracle can be either hostile or benign. On one extreme, it can deliberately use a hostile distribution over ω_1^t which maximally slows down the solver, *i.e.*

$$\sup_{\omega_1,\ldots,\omega_t \text{ all satisfy (17) and (18)}} \mathbb{E}_{\omega_1^t}[f(\mathbf{w}_{t+1}) - f^*].$$

On the other extreme, the oracle can use a benign distribution to help the solver reduce ϵ_{t+1} as fast as possible. Well, maybe the most solver-friendly distribution is to set the noise to zero which recovers the deterministic oracle[1].

1.4.3 Parametric form of stochastic oracle

In many real problems, the objective function is restricted to a parametric class, and the oracle can randomly choose a "parameter". Similar to Agarwal et al. (2009), we assume that all objective functions f can be written as the expectation of a parametric class of basis functions: with slight abuse of notation,

$$f(\mathbf{w}) = \mathop{\mathbb{E}}_{\alpha \sim Q}[f(\mathbf{w};\alpha)].$$

That is, f is determined by the distribution Q of α. For example, in binary SVM the parameter α corresponds to the data point (X, Y), f corresponds to the hinge loss, and Q is the joint distribution on the feature-label space.

[1] This claim is quite obvious, but I don't know how to prove it.

The oracle can take a "different" distribution P (even over a different random variable) and return

$$\phi(\mathbf{w}, f; \omega) = f(\mathbf{w}; \omega), \quad \text{where} \quad \omega \sim P.$$

The choice of P must satisfy (17) and (18). In the oracle, ω can be taken as the data point too, and P be set to Q. This is an extreme case where the stochastic oracle uses a single data point. Sometimes, mini-batch is also used, *e.g.* each step the oracle uses three *iid* samples:

$$\phi(\mathbf{w}, f; \omega_1, \omega_2, \omega_3) = \tfrac{1}{3}\left(f(\mathbf{w}; \omega_1) + f(\mathbf{w}; \omega_2) + f(\mathbf{w}; \omega_3)\right), \quad \text{where} \quad \omega_1, \omega_2, \omega_3 \sim Q.$$

2. Existing Minimax Lower Bounds with Hostile Stochastic Oracle

The main result on the minimax lower bound for stochastic convex optimization was given by Agarwal et al. (2009). We give a self-contained description of their results in this section.

Assume

$$\mathcal{F} = \left\{ f : \mathbb{R}^d \mapsto \mathbb{R} : f \text{ strongly convex and 1-Lipschitz continuous} \right\}$$
$$\mathbb{S} = \left\{ S \subseteq \mathbb{R}^d : S \text{ is convex}, \|\mathbf{x} - \mathbf{y}\|_\infty \le 1 \text{ for all } \mathbf{x}, \mathbf{y} \in S \right\}.$$

They also assume that the set of first-order oracle $\mathcal{O}$ consists of all unbiased samples of f and ∇f with bounded variance

$$\begin{aligned} \mathcal{O} := \{\phi : (f, \mathbf{w}) \mapsto & (\phi_1(f, \mathbf{w}; \omega), \boldsymbol{\phi}_2(f, \mathbf{w}; \omega)) : \\ & \mathbb{E}_\omega[\phi_1(f, \mathbf{w}; \omega)] = f(\mathbf{w}), \text{ and } \mathbb{E}_\omega[\boldsymbol{\phi}_2(f, \mathbf{w}; \omega)] = \nabla f(\mathbf{w}); \\ & \exists L > 0, \ s.t. \ \mathrm{Var}_\omega[\phi_1(f, \mathbf{w}; \omega)] < L \text{ and } \mathrm{Var}_\omega[\boldsymbol{\phi}_2(f, \mathbf{w}; \omega)] \preceq LI.\} \end{aligned} \tag{19}$$

There is no assumption on $\mathbb{M}_t$. Their conclusion is

$$\sup_{\phi \in \mathcal{O}} \inf_{\mathcal{M}_t \in \mathbb{M}_t} \sup_{S \in \mathbb{S}} \sup_{f \in \mathcal{F}} \mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)] \ge c\frac{d}{t}, \tag{20}$$

where c is an absolute constant.

Consider the subset $\mathcal{V} \subseteq \{-1, +1\}^d$ such that all pairs of points in $\mathcal{V}$ have Hamming distance at least $d/4$. By standard argument (Matoušek, 2002), there exists such a $\mathcal{V}$ whose cardinality is at least $(2/\sqrt{e})^{d/2}$. Choose the basis function set as

$$f_i^+(\mathbf{w}) := (w_i + 1/2)^2, \quad \text{and} \quad f_i^-(\mathbf{w}) := (w_i - 1/2)^2, \qquad \text{for } i = 1, \ldots, d. \tag{21}$$

And define the function set

$$\mathcal{G}_\epsilon := \left\{ g_{\boldsymbol{\alpha}} := \frac{1}{d}\sum_{i=1}^d \left\{ \left(\frac{1}{2} + \alpha_i \epsilon\right) f_i^+(\mathbf{w}) + \left(\frac{1}{2} - \alpha_i \epsilon\right) f_i^-(\mathbf{w}) \right\} : \boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_d) \in \mathcal{V} \right\}. \tag{22}$$

It is easy to check that $\mathcal{G}_\epsilon \subseteq \mathcal{F}$ over the L_∞ ball of radius $1/2$.

The idea of proof can be summarized in the following steps:

1. Fix arbitrary $\epsilon > 0$ and construct $\mathcal{V}$ and $\mathcal{G}_\epsilon$.

2. Recognize that for any arbitrary method $\mathcal{M}_t$ and $\boldsymbol{\alpha} \in \mathcal{V}$ which induces an objective function $g_{\boldsymbol{\alpha}} \in \mathcal{G}_\epsilon$, the process of stochastic optimization with a properly designed oracle is equivalent to estimating $\boldsymbol{\alpha}$ with t samples. That is, once the optimizer finds an ϵ accurate minimizer for $g_{\boldsymbol{\alpha}}$ in expectation, then $\boldsymbol{\alpha}$ can be correctly identified with high probability. The converse may also hold, but is not needed for the proof.

3. On the one hand, by Fano's lemma (see details in Lemma 3 of (Yu, 1993)), for any $\mathcal{M}_t^*$ and its induced estimator, there exists an $\boldsymbol{\alpha}^* \in \mathcal{V}$ such that the probability of correctly estimating $\boldsymbol{\alpha}^*$ is at most $O(\frac{t\epsilon}{d})$. This step relies on the cardinality of $\mathcal{V}$ (exponential in d), and imposes some restriction on the form of oracle. See details below.

4. On the other hand, once $\mathcal{M}_t^*$ finds an ϵ accurate solution of $g_{\boldsymbol{\alpha}^*}$ in expectation, it must be able to correctly identify $\boldsymbol{\alpha}^*$ with probability at least $\frac{1}{3}$. Markov inequality is the key to translating the expectation into probability.

5. Combining 3) and 4), one derives $t \geq O\left(\frac{d}{\epsilon}\right)$.

The invocation of Fano's lemma results in the minimax nature of the lower bound (for any method there exists a tough problem *s.t.* blah). This lemma is applied in the following context. Consider two random variables (Z, X) where Z and X are valued in $\{1, \ldots, d\}$ and $\{0, 1\}$ respectively. Fixing $\boldsymbol{\alpha} \in \mathcal{V}$, the joint distribution is defined by first setting $P(Z = i) = 1/d$ for all i, and then letting $P(X = 1|Z = i) = \frac{1}{2} + \alpha_i \epsilon$. At each iteration t, an *iid* sample (Z_t, X_t) is drawn, and the task is to estimate $\boldsymbol{\alpha}$. Given any estimator and a fixed number of samples, Fano's lemma states that there exists an $\boldsymbol{\alpha} \in \mathcal{V}$ such that the probability of incorrectly identifying $\boldsymbol{\alpha}$ can be lower bounded. Since Fano requires a sample of (Z, X) at each step, the following construction of oracle is natural for the objective in (22) (treating i_t as Z and b_{i_t} as X):

- Pick an $i_t \in \{1, \ldots, d\}$ uniformly at random.
- Draw $b_{i_t} \in \{0, 1\}$ according to a Bernoulli distribution with parameter $\frac{1}{2} + \alpha_{i_t} \epsilon$.
- Return the value and subgradient of the function $b_{i_t} f_{i_t}^+(\mathbf{w}) + (1 - b_{i_t}) f_{i_t}^-(\mathbf{w})$.

By construction, the returned function value and gradient are unbiased samples of those of $g_{\boldsymbol{\alpha}}$; moreover, the variance of the effective "noise" is bounded independently of d as long as the Lipschitz constant is independent of d since the function values and gradients are bounded on S.

If no strong convexity can be assumed, then the lower bound can be improved to $\Omega\left(\frac{d}{\epsilon^2}\right)$. The definition of function set $\mathcal{G}_\epsilon$ in (22) keeps intact, and we just need to modify (21) into

$$f_i^+(\mathbf{w}) := \left| w_i + \frac{1}{2} \right|, \quad \text{and} \quad f_i^-(\mathbf{w}) := \left| w_i - \frac{1}{2} \right|, \qquad \text{for } i = 1, \ldots, d.$$

2.1 Limitations of the lower bound

One clear restriction of the above lower bound is that the oracle must be chosen in a way that admits the application of Fano's lemma. This means that no lower bound is proved for the following oracle:

- Pick an $i_t \in \{1, \ldots, d\}$ uniformly at random.
- Return the value and subgradient of the function $\left(\frac{1}{2} + \alpha_{i_t}\epsilon\right) f_{i_t}^{+}(\mathbf{w}) + \left(\frac{1}{2} - \alpha_{i_t}\epsilon\right) f_{i_t}^{-}(\mathbf{w})$.

Another restriction lies in (19) where the expectation of the outcome of the oracle gives the true gradient. It is well known that more informative oracles also exist, *e.g.* (15), and the stochastic version just requires that the expectation of the oracle match it. In the deterministic setting, the oracle in (15) can be utilized to achieve $O(1/t^2)$ rate of convergence. Whether the same rate can be achieved for stochastic oracle is still not clear.

Besides, the oracle only returns the first order gradient (with noise), which forces the optimizer to perform explicit updates. Lots of practical experiences show that implicit update is useful. By implicit oracle, we mean that when optimizing $f(\mathbf{w}) = \mathbb{E}_{\alpha \sim Q}[f(\mathbf{w}; \alpha)]$, the oracle first draws an α_{t+1} from Q, and then returns the *function* $f(\mathbf{w}; \alpha_{t+1})$ rather than merely its gradient at $\mathbf{w}_t$.[2] So the result of the oracle may be even independent of $\mathbf{w}_t$. More generally, α_{t+1} could be drawn from another distribution P_{t+1} which depends on $\mathbf{w}_t$, as long as some unbiasedness conditions are maintained asymptotically. In fact, implicit update is an extremely reasonable scenario: in SVM training, the oracle returns a training example which determines a hinge loss function, rather than just its subgradient at a particular $\mathbf{w}_t$.

3. A Maximin Lower Bound for Uniform Distribution Oracle

This section dismisses all the hope outlined in the above Section 2.1 regarding bypassing the lower bound in (20). In particular, we consider an oracle ϕ which responses to $\mathbf{w}_t$ and $f = \mathbb{E}_{\alpha \in Q}[f(\mathbf{w}; \alpha)]$ with α_{t+1} drawn randomly from Q, together with the function $f(\mathbf{w}; \alpha_{t+1})$. This is clearly more informative than the stochastic first-order oracle. We show a maximin lower bound which is also stronger than the minimax lower bound given by Agarwal et al. (2009) in (20). Formally, we show that

$$\sup_{f,S} \inf_{\mathcal{M}_t} \mathbb{E}_\phi[\epsilon(\mathcal{M}_t, f, S, \phi)] \geq c \frac{d}{t},$$

where c is an absolute constant. f, S will be given below, $\mathcal{M}_t$ needs to satisfy some very mild conditions which will also be detailed below.

Let us consider the following one dimensional optimization problem:

$$\min_w f(w) = \min_w \frac{1}{2} \mathop{\mathbb{E}}_{X \sim Q}[(X - w)^2], \tag{23}$$

[2] We admit that the words "explicit update" and "implicit update" are not really accurate, because in online learning its meaning is just slightly similar.

where X has a pdf or pmf Q. To see the exact minimizer, decompose $f(w)$ in the standard way:

$$f(w) = \frac{1}{2}\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] + \frac{1}{2}(\mathbb{E}[X] - w)^2 \geq \frac{1}{2}\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = f(\mathbb{E}[X]).$$

So $f(w)$ is minimized at $w = \mathbb{E}[X]$ with minimal value $f^* = \frac{1}{2}\mathrm{Var}[X]$.

At each iteration, the stochastic oracle can provide a noisy gradient and function value as

$$\begin{array}{rcl} f'(w_t; X_t) & = & w_t - X_t \\ f(w_t; X_t) & = & \frac{1}{2}(X_t - w_t)^2 \end{array}, \quad \text{where} \quad X_t \sim Q. \tag{24}$$

Here X_t corresponds to ω_t introduced in Section 1.4.2, and we use X_t now to emphasize the meaning of sample. Unbiasedness is clear: $\mathbb{E}_{X_t \sim Q}[f'(w_t; X_t)] = w_t - \mathbb{E}_{X \sim Q}[X] = \nabla f(w_t)$ and $\mathbb{E}_{X_t \sim Q}[f(w_t; X_t)] = f(w_t)$. We also assume $X_1, \ldots, X_t$ are independent, and denote the sequence as X_1^t. In the sequel, it will become evident that the exact form of oracle in (24) is not important. It only needs to return a sample X_t and the optimizer can make use of $X_1, \ldots, X_t$ in whatever fashion, *e.g.* compute (24).

To bound ϵ_{t+1} we again decompose the objective function following the standard procedure:

$$\begin{aligned} \mathop{\mathbb{E}}_{X_1^t}[f(w_{t+1})] &= \mathop{\mathbb{E}}_{X_1^t}\left[\frac{1}{2}\mathop{\mathbb{E}}_{X}[(X - w_{t+1})^2]\right] = \frac{1}{2}\mathop{\mathbb{E}}_{X_1^t}\left[\mathbb{E}[(X - \mathbb{E}[X])^2] + (\mathbb{E}[X] - w_{t+1})^2\right] \\ &= f^* + \frac{1}{2}\underbrace{\left(\mathbb{E}[X] - \mathop{\mathbb{E}}_{X_1^t}[w_{t+1}]\right)^2}_{\text{bias square}} + \frac{1}{2}\underbrace{\mathop{\mathbb{E}}_{X_1^t}\left[\left(\mathop{\mathbb{E}}_{X_1^t}[w_{t+1}] - w_{t+1}\right)^2\right]}_{\text{variance}}. \end{aligned} \tag{25}$$

Since w_{t+1} is a function of *iid* samples $X_1, \ldots, X_t$ of X, we can view w_{t+1} as an estimator of $\mathbb{E}[X]$. In this sense, (25) can be decomposed into the square bias and variance of the estimator. This perspective also allows us to apply well known lower bounds in statistical estimation.

Suppose Q is parameterized by its mean $\theta := \mathbb{E}_Q[X]$, *i.e.* $P(X) = p(X; \theta)$. We restrict $\theta \in [0, 1]$. In this sense, we will also write f as f_θ. Denote the bias as

$$b_t(\theta) := \mathbb{E}_{X_i \sim p(X;\theta)}[w_{t+1}] - \theta.$$

Naturally, we assume the estimator is asymptotically unbiased:

Assumption 1. *w_t is asymptotically unbiased for all θ: $\lim_{t \to \infty} b_t(\theta) = 0, \ \forall \theta$.*

It is our goal to apply the Cramer-Rao's inequality which lower bounds the variance

$$v_t(\theta) := \mathbb{E}_{X_1^t}\left[\left(\mathbb{E}_{X_1^t}[w_{t+1}] - w_{t+1}\right)^2\right], \quad \text{where} \quad X_i \sim p(X; \theta)$$

by

$$v_t(\theta) \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_{X_i \sim p(X;\theta)}[w_{t+1}]\right)^2}{t \cdot \mathbb{E}_{X \sim p(X;\theta)}\left[(\frac{\mathrm{d}}{\mathrm{d}\theta}\log p(X;\theta))^2\right]} = \frac{(1 + b_t'(\theta))^2}{t \cdot \underbrace{\mathbb{E}_{X \sim p(X;\theta)}\left[(\frac{\mathrm{d}}{\mathrm{d}\theta}\log p(X;\theta))^2\right]}_{:= C(\theta)}}. \tag{26}$$

The expectation in the denominator is independent of the estimator (optimizer), and can be treated as a constant $C(\theta)$. The numerator requires differentiability of $b_t(\theta)$:

Assumption 2. *For all t, $b_t(\theta)$ is differentiable in θ.*

With these two assumptions, it is straightforward to show that the gap cannot be reduced faster than $O\left(\frac{1}{t}\right)$ uniformly over all θ.

Theorem 1. *Suppose Assumption 1 and 2 hold. Also assume that for any $\bar{\theta}$ there exists an open neighborhood in which $C(\theta)$ is upper bounded. Then there cannot exist a constant $c > 0$ and a superlinear function $g(t)$ (i.e., $\lim_{t\to\infty} \frac{g(t)}{t} = \infty$) such that*

$$\epsilon_{t+1} := \mathop{\mathbb{E}}_{X_i \sim p(X;\theta)} [f_\theta(w_{t+1})] - f_\theta^* \leq \frac{c}{g(t)} \quad \textit{for all } t \textit{ and } \theta. \tag{27}$$

Proof. Pick arbitrary $\theta_1 < \theta_3$. Then there exists $\theta_2 \in (\theta_1, \theta_3)$ such that $C(\theta)$ is upper bounded in $[\theta_1, \theta_2]$. Let $C < \infty$ be an upper bound. By (25) and (26), we have

$$\mathop{\mathbb{E}}_{X_i \sim p(X;\theta)} [f_\theta(w_{t+1})] - f_\theta^* = \frac{1}{2} b_t^2(\theta) + \frac{1}{2} v_t(\theta) \geq \frac{1}{2} b_t^2(\theta) + \frac{(1 + b_t'(\theta))^2}{2tC(\theta)}.$$

If (27) holds, then for all $\theta \in [\theta_1, \theta_2]$, we have

$$\frac{1}{2} b_t^2(\theta) + \frac{(1 + b_t'(\theta))^2}{2tC(\theta)} \leq \frac{c}{g(t)} \quad \Rightarrow \quad \frac{1}{2} b_t^2(\theta) + \frac{(1 + b_t'(\theta))^2}{2tC} \leq \frac{c}{g(t)} \quad \Rightarrow \quad b_t'(\theta) \leq -1 + \frac{c'}{\sqrt{g(t)/t}},$$

where c' is a constant independent of t and θ. Let $\gamma := \frac{\theta_2 - \theta_1}{4}$. There must exist a T such that for all $t > T$ we have

$$|b_t(\theta_1)| < \gamma, \quad |b_t(\theta_2)| < \gamma \tag{28}$$

$$b_t'(\theta) < -\frac{1}{2} \quad \forall\ \theta \in [\theta_1, \theta_2]. \tag{29}$$

By Role's mid-value theorem, there exists $\bar{\theta} \in [\theta_1, \theta_2]$, such that

$$|b_t(\theta_1) - b_t(\theta_2)| = \left|b_t'(\bar{\theta})\right| \cdot |\theta_1 - \theta_2| \, .$$

The left-hand side is less than $2\gamma = \frac{1}{2}(\theta_2 - \theta_1)$ by (28), while the right-hand side is greater than $\frac{1}{2}(\theta_2 - \theta_1)$ by (29). Contradiction. □

Theorem 1 is weak in that the assertion (27) requires the existence of a common c and $g(t)$ for all θ. To remove this uniform requirement and tighten the lower bound, we can request some more restrictive conditions on $b_t'(\theta)$. One practical assumption is the uniform boundedness which allows us to apply the dominated convergence theorem.

Assumption 3. *There exists an integrable function $h(\theta)$ such that $|b_t'(\theta)| \leq h(\theta)$ for all t.*

Theorem 2. *Suppose Assumption 1, 2 and 3 hold. Then the following is impossible: for all θ, there exists a constant $c_\theta > 0$ and a superlinear function $g_\theta(t)$ such that*

$$\mathop{\mathbb{E}}_{X_i \sim p(X;\theta)} [f_\theta(w_{t+1})] - f_\theta^* \leq \frac{c_\theta}{g_\theta(t)} \quad \textit{for all } t. \tag{30}$$

Proof. For any θ, by (25) and (26), we have

$$\mathop{\mathbb{E}}_{X_i \sim p(X;\theta)} [f_\theta(w_{t+1})] - f_\theta^* = \frac{1}{2} b_t^2(\theta) + \frac{1}{2} v_t(\theta) \geq \frac{1}{2} b_t^2(\theta) + \frac{(1 + b_t'(\theta))^2}{2tC(\theta)}.$$

If (30) holds, then

$$\frac{1}{2} b_t^2(\theta) + \frac{(1 + b_t'(\theta))^2}{2tC(\theta)} \leq \frac{c_\theta}{f_\theta(t)} \quad \forall\ t.$$

Thus, $\lim_{t\to\infty} b_t(\theta) = 0$ and $\lim_{t\to\infty} b_t'(\theta) = -1$ for all θ.

Arbitrarily pick two different θ_1 and θ_2. Thanks to Assumption 3, the dominated convergence theorem is applicable:

$$\lim_{t\to\infty} \int_{\theta_1}^{\theta_2} b_t'(\theta) \mathrm{d}\theta = \int_{\theta_1}^{\theta_2} \lim_{t\to\infty} b_t'(\theta) \mathrm{d}\theta. \tag{31}$$

Since $\lim_{t\to\infty} b_t'(\theta) = -1$ for all θ, so the right-hand side equals $\theta_1 - \theta_2 \neq 0$. But the left-hand side equals $\lim_{t\to\infty}(b_t(\theta_2) - b_t(\theta_1)) = 0 - 0 = 0$. Contradiction. □

3.1 Discussions on the assumptions

It seems that Assumption 3 is already the weakest condition needed for Theorem 2 to hold. A counter-example can be constructed to show that Theorem 2 may be broken if $b_t'(\theta)$ cannot be dominated by a fixed function. Simply set $b_t'(\theta)$ on $\theta \in [0, 1]$ as (for large enough t):

$$\begin{array}{rcll} b_t'(\theta) & = & -1, & \forall \theta \in [0, t^{-1}], \\ b_t'(\theta) & = & t, & \forall \theta \in (t^{-1}, t^{-1} + t^{-2}], \\ b_t'(\theta) & = & -1, & \forall \theta \in (t^{-1} + t^{-2}, t^{-1} + t^{-2} + t^{-1}], \\ b_t'(\theta) & = & t, & \forall \theta \in (t^{-1} + t^{-2} + t^{-1}, t^{-1} + t^{-2} + t^{-1} + t^{-2}], \\ & & \ldots & \end{array}$$

Denote $\Theta_t := \{\theta : b_t'(\theta) = -1\}$, and $\Theta_t' := \{\theta : b_t'(\theta) = t\}$. Then the measure/size of Θ_t is t times of Θ_t'. When t tends to infinity, "the number of" θ satisfying $b_t'(\theta) \neq -1$ approaches 0. One can show that for any θ, it is impossible to have infinitely many t such that $b_t'(\theta) \neq -1$. Integrating $b_t'(\theta)$ from 0 to any θ, we get that $b_t(\theta)$ lies in $[0, t^{-1}]$ (*i.e.*, uniformly approaches 0). Therefore, our construction is compatible with

$$\frac{1}{2} b_t^2(\theta) + \frac{(1 + b_t'(\theta))^2}{2C(\theta)t} \leq \frac{c_\theta}{f_\theta(t)} \quad \forall\ \theta, t.$$

To satisfy Assumption 2 and 3, various conditions can be imposed on the distribution $p(X;\theta)$, and the estimator w_{t+1}. For example, when $p(X;\theta)$ is an exponential family of distributions, Assumption 2 is satisfied for any measurable estimator w_{t+1}. Assumption 3 appears harder to check. A simple upper bound is a constant. Note that the boundedness can be guaranteed by the Lipschitz continuity of $b'_t(\theta)$ on a bounded domain of θ:

Assumption 4. *There exists an $L > 0$ such that for all t, the gradient of the bias with respect to θ, $b'_t(\theta)$, is Lipschitz continuous with constant L.*

3.2 Why the lower bound has not been bypassed

In Section 2.1, we outlined three possible ways to bypass the lower bound of (Agarwal et al., 2009). Now it is fairly easy to see that none of them survives the above counter-example. First, the oracle does return a sampled *function* (via a sample X_t) instead of just a subgradient. So it already accommodates implicit updates. Second, we have already used a most straightforward stochastic mechanism in the oracle (the first paragraph of Section 3). Finally, any other fancy oracle such as (15) is ultimately a function of the sample X_t. So it still falls in our framework for general estimators.

Acknowledgements

The authors thank Prof Arkadi Nemirovski for very helpful comments.

References

Jacob Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proc. Annual Conf. Computational Learning Theory*, 2008.

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Neural Information Processing Systems*, 2009.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In J. C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *NIPS*. MIT Press, 2007. http://books.nips.cc/papers/files/nips20/NIPS2007_0726.pdf.

Jiří Matoušek. *Lectures on Discrete Geometry*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2002.

Arkadi Nemirovski and D Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley and Sons, 1983.

Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005a.

Yurii Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM J. on Optimization*, 16(1):235–249, 2005b. ISSN 1052-6234.

Bin Yu. *Assouad, Fano and Le Cam.* Springer-Verlag, 1993.