Statistical machine learning and convex optimization	**2016**
Lecture 4 — March 24th	
Lecturer: Francis Bach	*Scribe: Yaheng Cui, Mounia Hamidouche*

4.1 Motivation

We have explored a variety of optimization methods in class (e.g., Newton's method, Gradient descent), which apply to differentiable functions. However, many functions that arise in practice are convex but non-differentiable at certain places, then it seems rather natural to replace gradients by subgradients.

In a lot of applications, the exact subgradient could be difficult to calculate because of errors in measurements, uncertainty in the data or because of errors in Monte Carlo evaluation of a function defined as an expected value. However, it is usually possible to get a noisy (unbiased) estimate to the subgradient. In this case, we can use the noisy estimate as the true value in the subgradient method, which is called the *stochastic subgradient method*.

Our main objective is to minimize a function f defined on $\mathbb{R}^d$ given only unbiased estimates $f_n'(\theta_n)$ of it's gradient $f'(\theta_n)$ at certain point $\theta_n \in \mathbb{R}^d$.

4.2 Stochastic approximation and machine learning

The minimization of an objective function which is only available through unbiased estimates of the function values or its gradients is a key methodological problem in many disciplines. Its analysis has been attacked mainly in two communities : *stochastic approximation and machine learning*.

4.2.1 Stochastic approximation

Stochastic approximation methods are a family of iterative stochastic algorithms that attempt to find zeroes or extremas of functions which cannot be computed directly, but only estimated via noisy observations. The structure of the algorithm for broader applicability beyond convex optimization is to generate iterates of the form :

$$\theta_n = \theta_{n-1} - \gamma_n h_n(\theta_{n-1}) \; with \;\; \mathbb{E}[h_n(\theta_{n-1})|\theta_{n-1}] = h(\theta_{n-1}).$$

For more information on the subject, see [6] and lecture 3.

4.2.2 Machine learning

We make the following assumptions:

- $\mathcal{H}$ is a d-dimensional Euclidean space, with $d \geq 1$.

- The observations $(x_n, y_n) \in \mathcal{H} \times \{-1, 1\}$ are independent and identically distributed.
- We considere $f_n(\theta) = \ell(y_n, \theta^T \phi(x_n))$ loss for a single pair of observations.
- Convex optimization problems coming from supervised machine learning are typically of the form $f(\theta) = \mathbb{E} f_n(\theta) = \mathbb{E} l(y_n, \theta^T \phi(x_n))$, which is the generalization error.
- Expected gradient: $f'(\theta) = \mathbb{E} f_n'(\theta) = \mathbb{E}\{\ell'(y_n, \theta^T \phi(x_n))\}$.

4.3 Relationship to online learning

Some stochastic optimization methods can optimize any convex function f over a convex domain given access only to unbiased estimates of $f's$ gradients. This feature makes it very useful for learning problems. Our goal is to minimize generalization error of θ $f(\theta) = \mathbb{E}_z \ell(\theta, z)$, using the gradients of single i.i.d. observation.

The goal of a learning system consists of finding the minimum of a function $\hat{f}(\theta)$ named the expected risk function.

4.3.1 Batch learning

The expected risk function $\hat{f}(\theta)$ cannot be minimized directly because the ground truth distribution is unknown. It is however possible to compute an approximation of $\hat{f}(\theta)$ by simply using a finite training set of independent observations $z_1, ..., z_n$.

- Given a finite set of observation: $z_1, ..., z_n$.
- Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^{n} \ell(\theta, z_k)$.
- $\hat{\theta}$ the minimizer of $\hat{f}(\theta)$ over a certain class Θ.
- Generalization bound using uniform concentration results (see lecture 1).

4.3.2 Online learning

Online machine learning is a method of machine learning in which data becomes available in a sequential order and is used to update our best predictor for future data at each step, as opposed to batch learning techniques which generate the best predictor by learning on the entire training data set at once.

- Update $\hat{\theta_n}$ after each new (potentially adversarial) observation z_n.
- Cumulative loss: $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^{n} \ell(\hat{\theta}_{k-1}, z_k)$.
- Online to batch through averaging [7].

4.4 Convex stochastic approximation

4.4.1 Some key propreries of f and f_n

Some key properties will be assumed for f and/or f_n :
- Smoothness: f B-Lipschitz continuous, f' L-Lipschitz continuous.
- f μ-strongly convex.

4.4.2 Some key algorithms

The main algorithms which have emerged (and which we will study) are stochastic gradient descent (a.k.a. Robbins-Monro algorithm) for which the structure is to generate iterates of the form:

$$\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1}).$$

We consider well as a simple modification where iterates are averaged (a.k.a. Polyak-Ruppert averaging):

$$\bar{\theta_n} = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k.$$

For the classical learning rate sequence : $\gamma_n = Cn^{-\alpha}$, for $\alpha \in [0,1]$ and C to be determined.

4.5 Stochastic subgradient "descent" method

We consider the unconstrained minimization of a function f that satisfy the following assumptions:

- f_n convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leq D\}$.
- (f_n) i.i.d. function such that $\mathbb{E}f_n = f$.
- θ_* global optimum of f on $\mathcal{C} = \{\|\theta\|_2 \leq D\}$.

Here we need to be a bit careful if we want to analyze this scheme under the three assumptions. Indeed we have a control on the size of the subgradients only in a ball of radius D. Furthermore we also know that the minimizer of the function lies in this ball. Thus it makes sense to enforce that if we leave the ball, then we first project back the point to the ball before taking another gradient step. This gives the (projected) subgradient "descent". (note that the method may occasionally go up).

The update of the iterate θ is as follows:

$$\theta_n = \Pi_D\left(\theta_{n-1} - \frac{2D}{B\sqrt{n}}f_n'(\theta_{n-1})\right).$$

The following elementary result gives a rate of convergence for the subgradient method.
Theorem :

$$\mathbb{E}f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \leq \frac{2DB}{\sqrt{n}}.$$

Proof:

- $\mathbb{F}_n$:information up to time n.
- $\| f_n'(\theta \|_2 \leq B$ and $\| \theta \|_2 \leq D$, unbiased gradients/function $\mathbb{E}(f_n \mid \mathcal{F}_{n-1}) = f$.

By contractivity of the projection we get:

$$\| \theta_n - \theta_* \|_2^2 \leq \| \theta_{n-1} - \theta_* - \gamma_n f_n'(\theta_{n-1}) \|_2^2 .$$

Then because $\| f_n'(\theta_{n-1}) \|_2 \leq B$ we have:

$$\| \theta_n - \theta_* \|_2^2 \leq \| \theta_{n-1} - \theta_* \|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^T f_n'(\theta_{n-1}).$$

$$\mathbb{E}[\| \theta_n - \theta_* \|_2^2 \,|\mathcal{F}_{n-1}] \leq \| \theta_{n-1} - \theta_* \|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^T f'(\theta_{n-1}).$$

From the subgradient proprety we have:

$$\mathbb{E}[\| \theta_n - \theta_* \|_2^2 \,|\mathcal{F}_{n-1}] \leq \| \theta_{n-1} - \theta_* \|_2^2 + B^2\gamma_n^2 - 2\gamma_n[f(\theta_{n-1}) - f(\theta_*)].$$

$$\mathbb{E} \| \theta_n - \theta_* \|_2^2 \leq \mathbb{E} \| \theta_{n-1} - \theta_* \|_2^2 + B^2\gamma_n^2 - 2\gamma_n[\mathbb{E}f(\theta_{n-1}) - f(\theta_*)].$$

Leading to:

$$\mathbb{E}f(\theta_{n-1}) - f(\theta_*) \leq \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n}[\mathbb{E} \| \theta_{n-1} - \theta_* \|_2^2 - \mathbb{E} \| \theta_n - \theta_* \|_2^2] \quad (*).$$

Starting from $(*)$ we have :

$$\sum_{u=1}^{n}[\mathbb{E}f(\theta_{u-1}) - f(\theta_*)] \leq \sum_{u=1}^{n}\frac{B^2\gamma_u}{2} + \sum_{u=1}^{n}\frac{1}{2\gamma_u}[\mathbb{E} \| \theta_{u-1} - \theta_* \|_2^2 - \mathbb{E} \| \theta_u - \theta_* \|_2^2].$$

$$\leq \sum_{u=1}^{n}\frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_n} \leq 2DB\sqrt{n} \quad \text{with} \quad \gamma_n = \frac{2D}{B\sqrt{n}}.$$

By using convexity we get the following result: $\mathbb{E}f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \leq \frac{2DB}{\sqrt{n}}$.

Remarks:

- Minimax rate (Nemirovsky and Yudin, 1983; Agarwal et al., 2012).

- Running-time complexity: $O(dn)$ after n iterations. Note the difference with the $O(dn^2)$ complexity of minimizing the empirical risk with the subgradient method.

4.6 Stochastic subgradient method: Extension to online learning

Assume different and arbitrary functions $f_n : \mathbb{R}^d \Longrightarrow \mathbb{R}$.

-Observations of $f_n'(\theta_{n-1}) + \epsilon_n$.

-with $\mathbb{E}(\epsilon|\mathcal{F}_{n-1}) = 0$ and $\|f_n'(\theta_{n-1}) + \epsilon_n\| \leq B$ Almost surely.

The performance criterion : (normalized) regret is defined as follows:

$$\frac{1}{n}\sum_{u=1}^{n} f_i(\theta_{i-1}) - \inf_{\|\theta\|_2 \leq D} \frac{1}{n}\sum_{i=1}^{n} f_i(\theta).$$

Remarks

- The regret is ofter not normalized.
- May not be non-negative (typically is).

Theorem :

The iteration of the algorithm is given as

$$\theta = \Pi_D(\theta_{n-1} - \gamma_n(f_n'(\theta_{n-1}) + \epsilon_n)).$$

for which the learning rate sequence is $\gamma_n = \frac{2D}{B\sqrt{n}}$.

Proof : (essentially the same as for stochastic approximation)

- $\mathcal{F}_n$: information up to time n.
- θ an arbitrary point such that $\| \theta \| \leq D$.
- $\| f_n'(\theta_{n-1}) + \epsilon_n \|_2 \leq B$ and $\| \theta \|_2 \leq D$, unbiased gradients $\mathbb{E}(\epsilon_n | \mathcal{F}_{n-1}) = 0$.

By contractivity of projections we have:

$$\| \theta_n - \theta \|_2^2 \leq \| \theta_{n-1} - \theta - \gamma_n(f_n'(\theta_{n-1}) + \epsilon_n) \|_2^2 .$$

Since $\| f_n'(\theta_{n-1}) + \epsilon_n \|_2$ then :

$$\| \theta_n - \theta \|_2^2 \leq \| \theta_{n-1} - \theta \|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta)^T(f_n'(\theta_{n-1}) + \epsilon_n)$$

.

$$\mathbb{E}[\| \theta_n - \theta \|_2^2 | \mathcal{F}_{n-1}] \leq \| \theta_{n-1} - \theta \|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta)^T f_n'(\theta_{n-1}).$$

From subgradient proprety we have:

$$\mathbb{E}[\| \theta_n - \theta \|_2^2 | \mathcal{F}_{n-1}] \leq \| \theta_{n-1} - \theta \|_2^2 + B^2\gamma_n^2 - 2\gamma_n[f_n(\theta_{n-1}) - f_n(\theta)]$$

$$\mathbb{E} \| \theta_n - \theta \|_2^2 \leq \mathbb{E} \| \theta_{n-1} - \theta \|_2^2 + B^2\gamma_n^2 - 2\gamma_n[\mathbb{E}f_n(\theta_{n-1}) - f_n(\theta)].$$

Leading to:

$$\mathbb{E}f_n(\theta_{n-1}) - f_n(\theta) \leq \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n}[\mathbb{E} \| \theta_{n-1} - \theta \|_2^2 - \mathbb{E} \| \theta_n - \theta \|_2^2] \ (*).$$

Starting from $(*)$, we have :

$$\sum_{u=1}^{n} [\mathbb{E}f_u(\theta_{u-1}) - f_u(\theta)] \leq \sum_{u=1}^{n} \frac{B^2\gamma_u}{2} + \sum_{u=1}^{n} \frac{1}{2\gamma_u}[\mathbb{E} \| \theta_{u-1} - \theta \|_2^2 - \mathbb{E} \| \theta_u - \theta \|_2^2].$$

Finally, for any θ such that $\| \theta \| \leq D$:

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}f_k(\theta_{k-1}) - \frac{1}{n}\sum_{k=1}^{n}f_k(\theta) \leq \frac{2DB}{\sqrt{n}}.$$

Online to batch conversion: assuming convexity, we can get back the results from stochastic approximation by using Jensen's inequality.

4.7 Stochastic subgradient descent -strong convexity 1-

Now we will talk about another property of convex functions that can significantly speed-up the convergence : strong convexity. We say that $f : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex if it satisfies:

$$f(x) - f(y) \leq \bigtriangledown f(x)^T(x-y) - \frac{\alpha}{2}\| x-y \|^2 .$$

Of course this definition does not require differentiability of the function f.

In this section we investigate the setting where f is *strongly convex* but potentially *non-smooth*. As we have already seen in a previous section, in the case of non-smooth functions we have to project back on the set where we control the norm of the gradients.

We consider the unconstrained minimization of a function f that satisfies the following requirements:

- f_n convex and B-Lipschitz-continuous.
- (f_n) i.i.d function such that $\mathbb{E}f_n = f$.
- f μ-srongly convex on $\{\|\theta\|_2 \leq D\}$. Note that we do not assume f_n to be strongly-convex.
- θ_* global optimum of f over $\{\|\theta\|_2 \leq D\}$.

Theorem:

We consider the projected subgradient descent algorithm with time-varying step size, that is:

$$\theta_n = \Pi_D\left(\theta_{n-1} - \frac{2}{\mu(n+1)}f_n'(\theta_{n-1})\right),$$

with the bound

$$\mathbb{E}f\left(\frac{2}{n(n+1)}\sum_{k=1}^{n}k\theta_{k-1}\right) - f(\theta_*) \leq \frac{2B^2}{\mu(n+1)}.$$

Proof:

As in the previous cases, by contractivity of projection we have:

$$\| \theta_n - \theta_* \|_2^2 \leq \| \theta_{n-1} - \theta_* - \gamma_n f_n'(\theta_{t-1}) \|_2^2 .$$

Because $\| f_n'(\theta_{t-1}) \|_2 \leq B$ we have:

$$\| \theta_n - \theta_* \|_2^2 \leq \| \theta_{n-1} - \theta_* \|_2^2 \| + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^T f_n'(\theta_{t-1}),$$

$$\mathbb{E}(.|\mathcal{F}_{n-1}) \leq \| \theta_{n-1} - \theta_* \|_2^2 \| + [f(\theta_{n-1} - f(\theta_*)) + \frac{\mu}{2} \| \theta_{n-1} - \theta_* \|_2^2].$$

Therefore this leads to :

$$\begin{aligned} \mathbb{E}f(\theta_{n-1}) - f(\theta_*) &\leq \tfrac{B^2\gamma_n}{2} + \tfrac{1}{2}[\tfrac{1}{\gamma_n} - \mu] \| \theta_{n-1} - \theta_* \|_2^2 - \tfrac{1}{2\gamma_n} \| \theta_n - \theta_* \|_2^2 . \\ &\leq \tfrac{B^2}{\mu(n+1)} + \tfrac{\mu}{2}[\tfrac{n-1}{2}] \| \theta_{n-1} - \theta_* \|_2^2 - \tfrac{\mu(n+1)}{4} \| \theta_n - \theta_* \|_2^2 . \end{aligned}$$

From $\mathbb{E}f(\theta_{n-1}) - f(\theta_*) \leq \tfrac{B^2}{\mu(n+1)} + \tfrac{\mu}{2}[\tfrac{n-1}{2}]\mathbb{E} \| \theta_{n-1} - \theta_* \|_2^2 - \tfrac{\mu(n+1)}{4}\mathbb{E} \| \theta_n - \theta_* \|_2^2 .$

$$\sum_{u=1}^{n} u[\mathbb{E}f(\theta_{u-1}) - f(\theta_*)] \leq \sum_{u=1}^{n} \frac{B^2 u}{\mu(u+1)} + \frac{1}{4}\sum_{u=1}^{n} u(u-1)\mathbb{E} \| \theta_{u-1} - \theta_* \|_2^2 - u(u+1)\mathbb{E} \| \theta_u - \theta_* \|.$$

$$\leq \sum_{u=1}^{n} \tfrac{B^2 u}{\mu(u+1} + \tfrac{1}{4}\sum_{u=1}^{n} u(u-1)\mathbb{E} \| \theta_n - \theta_* \|_2^2 \leq \tfrac{B^2 n}{\mu}.$$

• Using convexity :

$$\mathbb{E}f\left(\tfrac{2}{n(n+1)} \sum_{u=1}^{n} u\theta_{u-1}\right) - g(\theta_*) \leq \tfrac{2B^2}{n+1}.$$

Remarks:

- "Same " proof than deterministic case (Lacoste-Julien et al., 2012)
- This the minimax rate (Nemirovsky and Yudin, 1983; Agarwal et al., 2012).

4.8 Stochasic subgradient descent -strong convexity 2-

• Assumptions: we consider an unconstrained regularized problem.

- f_n convex and B-Lipschitz-continuous.
- (f_n) i.i.d. dunctions such that $\mathbb{E}f_n = f$.
- θ_* global optimum of $g = f + \frac{\mu}{2} \| . \|_2^2$.
- No compactness assumption - no projections (note the impossibility of having a strongly convex Lipschitz-continuous function on $\mathbb{R}^d$)

- Algorithm:

$$\theta_n = \theta_{n-1} - \tfrac{2}{\mu(n+1)} g'_n(\theta_{n-1}) = \theta_{n-1} - \tfrac{2}{\mu(n+1)}[f'_n(\theta_{n-1}) + \mu\theta_{n-1}]$$

- Bound:

$$\mathbb{E}g\left(\frac{2}{n(n+1)}\sum_{k=1}^{n} k\theta_{k-1}\right) - g(\theta_*) \leq \frac{2B^2}{\mu(n+1)}.$$

- Minimax convergence rate.

4.9 Strong convexity -Proof with $\log n$ factor-

Theorem:

Under the assumptions $\| f'_n(\theta) \|_2 \leq B$, $\| \theta \|_2 \leq D$ and $\mu - strong$ convexity of f, the iteration of the algorithm is given as

$$\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f'_n(\theta_{t-1}) \; with \; \gamma_n = \frac{1}{\mu n}$$

Proof:

by contractivity of projections we get:

$$\| \theta_n - \theta_* \|_2^2 \leq \| \theta_{n-1} - \theta_* - \gamma_n f'_n(\theta_{t-1}) \|_2^2$$

Since $\| f'_n(\theta_{t-1}) \|_2 \leq B$ then:

$$\| \theta_n - \theta_* \|_2^2 \leq \| \theta_{n-1} - \theta_* \|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^T f'_n(\theta_{t-1})$$

Using the proprety of subgradient and strong convexity:

$$\mathbb{E}(.|\mathcal{F}_{n-1}) \leq \| \theta_{n-1} - \theta_* \|_2^2 + B^2\gamma_n^2 - 2\gamma[f(\theta_{n-1}) - f(\theta_*) + \frac{\mu}{2} \| \theta_{n-1} - \theta_* \|_2^2]$$

Leading to :

$$\begin{aligned}\mathbb{E}f(\theta_{n-1}) - f(\theta_*) &\leq \frac{B^2\gamma_n}{2} + \frac{1}{2}[\frac{1}{\gamma_n} - \mu] \| \theta_{n-1} - \theta_* \|_2^2 - \frac{1}{2\gamma_2} \| \theta_n - \theta_* \|_2^2 \\ &\leq \tfrac{B^2\gamma_n}{2\mu n} + \tfrac{\mu}{2}[n-1] \| \theta_{n-1} - \theta_* \|_2^2 - \tfrac{n\mu}{2} \| \theta_n - \theta_* \|_2^2 \; (\bigstar).\end{aligned}$$

From ($\bigstar$):

$$\sum_{u=1}^{n} [\mathbb{E}f(\theta_{u-1}) - f(\theta_*)] \leq \sum_{u=1}^{n} \frac{B^2}{2n\mu} + \frac{1}{2}\sum_{u=1}^{n} [(u-1)\mathbb{E} \| \theta_n - \theta_* \|_2^2 - u\mathbb{E} \| \theta_u - \theta_* \|_2^2]$$

$$\leq \frac{B^2 logn}{2\mu} + \frac{1}{2}[0 - n\mathbb{E} \parallel \theta_n - \theta_* \parallel_2^2] \leq \frac{B^2 \log n}{2\mu}.$$

Using convexity : $\mathbb{E}f\left(\frac{1}{n}\sum_{u=1}^{n}\theta_{u-1}\right) - f(\theta_* \leq \frac{B^2 \log n}{2n\mu}$.

4.9.1 Relationship to online learning

Uniform averaging allows to get a bound for online learning.

For all θ :

$$\frac{1}{n}\sum_{i=1}^{n} f_i(\theta_{i-1}) - \frac{1}{n}\sum_{i=1}^{n} f_i(\theta) \leq \frac{B^2 \log n}{2n\mu}$$

Note that the $\log n$ term is not optimal; see Hazan and Kale (2012).

4.10 Beyond convergence in expectation

▶ **Typical result :** $Ef\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \leq \frac{2DB}{\sqrt{n}}$.

-Obtained with simple conditioning arguments.

▶ **High probability bounds**

-Markov inequality: $\mathbb{P}\left(f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_* \geq \epsilon\right) \leq \frac{2DB}{\sqrt{n}\epsilon}$.

-Concentration inequality (Nemirovski et al., 2009; Nesterov, 2009)

$$\mathbb{P}\left(f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \geq \frac{2DB}{\sqrt{n}}(2+4t)\right) \leq 2\exp(-t^2).$$

▶ See also Bach (2013) for logistic regression.

4.10.1 Stochastic subgradient method - high probability

We note that for a deterministic problem with extremely large scale ,using randomness may makes it easier or tractable. Then it will be crucial to derive some sort of probabilistic control of the algorithm and obtain some sort of "with high probability" convergence results.

Consider stochastic subgradient method with iteration:

$$\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f_n'(\theta_{n-1}))$$

where $\gamma_n = \frac{2D}{B\sqrt{n}}$, let $\mathcal{F}_n$ be the information from time 1 up to time n , $\|f_n'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$, the unbiased gradients/functions are given as:

$$\mathbb{E}(f_n|\mathcal{F}_{n-1}) = f$$

then we can have the following inequalities:

$$\begin{aligned}\|\theta_n-\theta_*\|_2^2 &\leqslant \|\theta_{n-1}-\theta_*-\gamma_n f_n'(\theta_{n-1})\|_2^2 \text{ by contractivity of projections}\\ &\leqslant \|\theta_{n-1}-\theta_*\|_2^2+B^2\gamma_n^2-2\gamma_n(\theta_{n-1}-\theta_*)^\top f_n'(\theta_{n-1}) \text{ because } \|f_n'(\theta_{n-1})\|_2\leqslant B\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\big[\|\theta_n-\theta_*\|_2^2|\mathcal{F}_{n-1}\big]&\leqslant \|\theta_{n-1}-\theta_*\|_2^2+B^2\gamma_n^2-2\gamma_n(\theta_{n-1}-\theta_*)^\top f'(\theta_{n-1})\\ &\leqslant \|\theta_{n-1}-\theta_*\|_2^2+B^2\gamma_n^2-2\gamma_n\big[f(\theta_{n-1})-f(\theta_*)\big] \text{ (subgradient property)}\end{aligned}$$

Set $Z_n=-2\gamma_n(\theta_{n-1}-\theta_*)^\top[f_n'(\theta_{n-1})-f'(\theta_{n-1})]$, then we can have that:

$$\begin{aligned}\|\theta_n-\theta_*\|_2^2 &\leqslant \|\theta_{n-1}-\theta_*\|_2^2+B^2\gamma_n^2-2\gamma_n\big[f(\theta_{n-1})-f(\theta_*)\big]+Z_n\\ f(\theta_{n-1})-f(\theta_*) &\leqslant \frac{1}{2\gamma_n}\big[\|\theta_{n-1}-\theta_*\|_2^2-\|\theta_n-\theta_*\|_2^2\big]+\frac{B^2\gamma_n}{2}+\frac{Z_n}{2\gamma_n}\end{aligned}$$

it follows the result that:

$$\begin{aligned}\sum_{u=1}^n\big[f(\theta_{u-1})-f(\theta_*)\big]&\leqslant \sum_{u=1}^n\frac{B^2\gamma_u}{2}+\sum_{u=1}^n\frac{1}{2\gamma_u}\big[\|\theta_{u-1}-\theta_*\|_2^2-\|\theta_u-\theta_*\|_2^2\big]+\sum_{u=1}^n\frac{Z_u}{2\gamma_u}\\ &\leqslant \sum_{u=1}^n\frac{B^2\gamma_u}{2}+\frac{4D^2}{2\gamma_n}+\sum_{u=1}^n\frac{Z_u}{2\gamma_u}\leqslant\frac{2DB}{\sqrt{n}}+\sum_{u=1}^n\frac{Z_u}{2\gamma_u} \text{ with } \gamma_n=\frac{2D}{B\sqrt{n}}\end{aligned}$$

The inequality above involves the analysis of the properties of

$$\sum_{u=1}^n\frac{Z_u}{2\gamma_u},$$

where $\mathbb{E}(Z_n|\mathcal{F}_{n-1})=0$ and $|Z_n|\leqslant 8\gamma_n DB$.

With the conditions of Z_n above, we also have $\mathbb{E}(\frac{Z_n}{2\gamma_n}|\mathcal{F}_{n-1})=0$ and $|\frac{Z_n}{2\gamma_n}|\leqslant 4DB$, if we take $X_n=\sum_{u=1}^n\frac{Z_u}{2\gamma_u}$, then we have $\{X_k:k=0,1,2,...\}$ is a martingale and $|X_k-X_{k-1}|\leq 4DB$, which allows also to use the Azuma–Hoeffding inequality who gives a concentration result for the values of martingales that have bounded differences which is :

$$\mathbb{P}\Big(\sum_{u=1}^n\frac{Z_u}{2\gamma_u}\geqslant t\sqrt{n}\cdot 4DB\Big)\leqslant\exp\Big(-\frac{t^2}{2}\Big).$$

4.11 Beyond stochastic gradient method

In machine learning, online algorithms operate by repetitively drawing random examples, one at a time, and adjusting the learning variables using simple calculations that are usually based on the single example only. The low computational complexity (per iteration) of online

algorithms is often associated with their slow convergence and low accuracy in solving the underlying optimization problems. But the combined low complexity and low accuracy, together with other tradeoffs in statistical learning theory, still make online algorithms favorite choices for solving large-scale learning problems. Nevertheless, traditional online algorithms, such as stochastic gradient descent, have limited capability of exploiting problem structure in solving regularized learning problems.

Some studies on the new class of online algorithms were developed by adding a proximal step, the regularized stochastic learning problems considered are of the following form:

$$\min_{\theta\in\mathbb{R}^d} f(\theta) + \Omega(\theta) = \mathbb{E} f_n(\theta) + \Omega(\theta)$$

and remplace the recursion $\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_n)$ by

$$\theta_n = \min_{\theta\in\mathbb{R}^d} \left\|\theta - \theta_{n-1} + \gamma_n f_n'(\theta_n)\right\|_2^2 + C\Omega(\theta)$$

Related work can be find by Xiao (2010); Hu et al. (2009). In addition, the simulation results by Ghadimi and Lan (2013) showing that the proposed algorithms have behavior similar to that of accelerated stochastic subgradient method and the primal-dual averaging method of Nesterov.

The related frameworks can be classed in two groups:

- Regularized dual averaging (Nesterov, 2009; Xiao, 2010) .
- Mirror descent (Nemirovski et al., 2009; Lan et al.,2012).

The following section will focus on The Mirror Descent Algorithm.

4.12 Mirror Descent

The key advantage of the subgradient algorithm is its simplicity, provided that projections can be easily computed, which is the case when the constraints set is described by simple sets, e.g., hyperplanes, balls, bound constraints, etc. Its main drawback is that it has a very slow rate of convergence.

The idea of the Mirror Descent algorithm is based on dealing with the structure of the Euclidean norm rather than with local behavior of the objective function in problem. Roughly speaking, the method originated from functional analytic arguments arising within the infinite dimensional setting, between primal and dual spaces. So we can see Mirror Descent as projected (stochastic) gradient descent adapted to Euclidean geometry which have the bound:

$$\frac{\max_{\theta,\theta'\in\Theta} \|\theta - \theta'\|_2 \cdot \max_{\theta\in\Theta} \|f'(\theta)\|_2}{\sqrt{n}}.$$

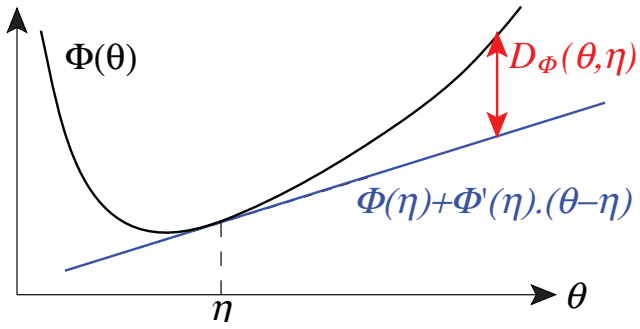
If we consider other norms instead of Euclidean norm in our model, the bound will have some changes, for example: natural bound on $\max_{\theta\in\Theta} \|f'(\theta)\|_\infty$ leads to $\sqrt{d}$ factor, and avoidable with mirror descent, which leads to factor $\sqrt{\log d}$. Some related work can be also find in Nemirovski et al. (2009); Lan et al. (2012).

4.12.1 Mirror descent set-up

For better understanding of Mirror descent, we should be familiar with the following notions:

- Objective function f is defined on domain $\mathcal{C}$;
- Arbitrary norm $\|\cdot\|$ with dual norm is defined as $\|s\|_* = \sup_{\|\theta\|\leqslant 1} \theta^\top s$;
- f is a B-Lipschitz-continuous function w.r.t. $\|\cdot\|$ such that: $\|f'(\theta)\|_* \leqslant B$;
- Given a strictly-convex function Φ, define the Bregman divergence

$$D_\Phi(\theta, \eta) = \Phi(\theta) - \Phi(\eta) - \Phi'(\eta)^\top(\theta - \eta)$$



4.12.2 Mirror map

Consider a strongly-convex function $\Phi : \mathcal{C}_\Phi \to \mathbb{R}$ such that

(a) the gradient Φ' takes all possible values in $\mathbb{R}^d$ which leads to a bijection from $\mathcal{C}_\Phi$ to $\mathbb{R}^d$;

(b) the gradient Φ' diverges on the boundary of $\mathcal{C}_\Phi$;

(c) $\mathcal{C}_\Phi$ contains the closure of the domain $\mathcal{C}$ of the optimization problem;

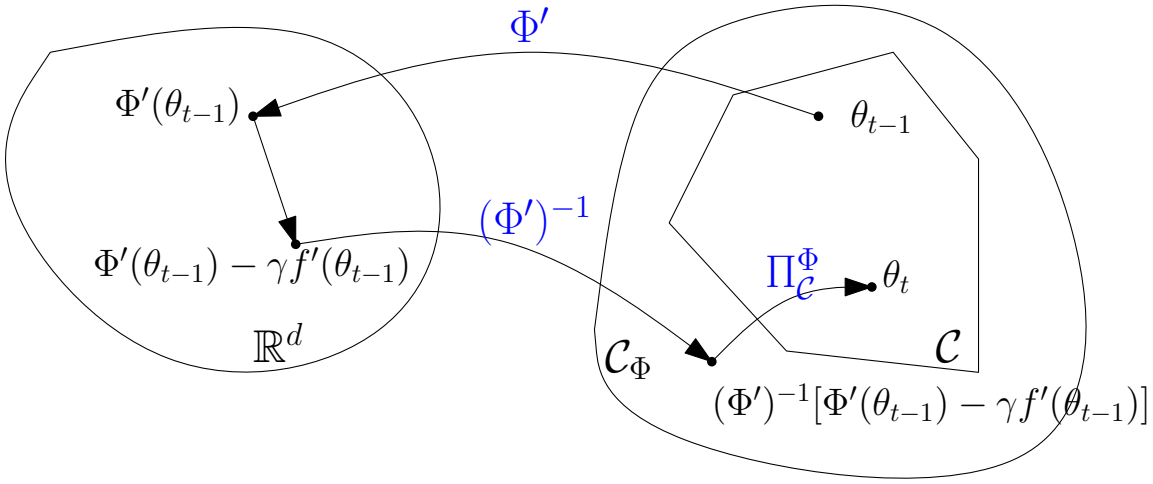
Then Bregman projection on $\mathcal{C}$ uniquely defined on $\mathcal{C}_\Phi$:

$$\begin{aligned}
\Pi_{\mathcal{C}}^{\Phi}(\theta) &= \arg\min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} D_\Phi(\eta, \theta) \\
&= \arg\min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} \Phi(\eta) - \Phi(\theta) - \Phi'(\theta)^\top(\eta - \theta) \\
&= \arg\min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} \Phi(\eta) - \Phi'(\theta)^\top \eta
\end{aligned}$$

4.12.3 Mirror descent

The iteration for Mirror descent is described as:

$$\theta_t = \Pi_{\mathcal{C}}^{\Phi}\big(\Phi'^{-1}\big[\Phi'(\theta_{t-1}) - \gamma f'(\theta_{t-1})\big]\big)$$



Convergence

Assume (a) $D^2 = \sup_{\theta\in\mathcal{C}} \Phi(\theta) - \inf_{\theta\in\mathcal{C}} \Phi(\theta)$, (b) Φ is α-strongly convex with respect to $\|\cdot\|$ and (c) f is B-Lipschitz-continuous wr.t. $\|\cdot\|$. Then with $\gamma = \frac{D}{B}\sqrt{\frac{2\alpha}{t}}$, we have:

$$f\left(\frac{1}{t}\sum_{u=1}^{t}\theta_u\right) \leqslant DB\sqrt{\frac{2}{\alpha t}}$$

Remark:

- See detailed proof in Bubeck (2015, p. 299).
- "Same" as subgradient method but allows stochastic gradients.

Proof (1)Define $\Phi'(\eta_t) = \Phi'(\theta_{t-1}) - \gamma f'(\theta_{t-1})$. We have

$$\begin{aligned} f(\theta_{t-1}) - f(\theta) &\leqslant f'(\theta_{t-1})^\top(\theta_{t-1} - \theta) = \frac{1}{\gamma}(\Phi'(\theta_{t-1}) - \Phi'(\eta_t))^\top(\theta_{t-1} - \theta) \\ &= \frac{1}{\gamma}\big[D_\Phi(\theta, \theta_{t-1}) + D_\Phi(\theta_{t-1}, \eta_t) - D_\Phi(\theta, \eta_t)\big] \end{aligned}$$

(2) By optimality of θ_t: $(\Phi'(\theta_t) - \Phi'(\eta_t))^\top(\theta_t - \theta) \leqslant 0$ which is equivalent to: $D_\Phi(\theta, \eta_t) \geqslant D_\Phi(\theta, \theta_t) + D_\Phi(\theta_t, \eta_t)$. Thus

$$\begin{aligned} D_\Phi(\theta_{t-1}, \eta_t) - D_\Phi(\theta_t, \eta_t) &= \Phi(\theta_{t-1}) - \Phi(\theta_t) - \Phi'(\eta_t)^\top(\theta_{t-1} - \theta_t) \\ &\leqslant (\Phi'(\theta_{t-1}) - \Phi'(\eta_t))^\top(\theta_{t-1} - \theta_t) - \frac{\alpha}{2}\|\theta_{t-1} - \theta_t\|^2 \\ &= \gamma f'(\theta_{t-1})^\top(\theta_{t-1} - \theta_t) - \frac{\alpha}{2}\|\theta_{t-1} - \theta_t\|^2 \\ &\leqslant \gamma B\|\theta_{t-1} - \theta_t\| - \frac{\alpha}{2}\|\theta_{t-1} - \theta_t\|^2 \leqslant \frac{(\gamma B)^2}{2\alpha} \end{aligned}$$

(3) Thus

$$\sum_{u=1}^{t} \big[f(\theta_{t-1}) - f(\theta)\big] \leqslant \frac{D_\Phi(\theta, \theta_0)}{\gamma} + \gamma \frac{L^2 t}{2\alpha}.$$

■

4.13 Minimax rates (Agarwal et al., 2012)

In order to measure the hardness of an optimization problem, we consider:

(a) A model of computation (i.e., algorithms): first-order oracle

- Queries a function f by obtaining $f(\theta_k)$ and $f'(\theta_k)$ with zero-mean bounded variance noise, for $k = 0, \dots, n-1$ and outputs θ_n

(b) A class of functions

- convex B-Lipschitz-continuous (w.r.t. ℓ_2-norm) on a compact convex set $\mathcal{C}$ containing an ℓ_∞-ball

(c) A performance measure

- for a given algorithm and function $\varepsilon_n(\text{algo}, f) = f(\theta_n) - \inf_{\theta \in \mathcal{C}} f(\theta)$
- for a given algorithm: $\sup_{\text{functions } f} \varepsilon_n(\text{algo}, f)$

We then define the Minimax performance as: $\inf_{\text{algo}} \sup_{\text{functions } f} \varepsilon_n(\text{algo}, f)$. We are going to prove the following:

- **Convex functions**: domain $\mathcal{C}$ that contains an ℓ_∞-ball of radius D

$$\inf_{\text{algo}} \sup_{\text{functions } f} \varepsilon(\text{algo}, f) \geqslant \text{ cst } \times \min\left\{ BD\sqrt{\frac{d}{n}}, BD \right\}$$

  - This implies the following bound for the ℓ_2-ball of radius D: $BD/\sqrt{n}$
  - The upper-bound is obtained through through stochastic subgradient: they match!

- **μ-strongly-convex functions**: domain $\mathcal{C}$ that contains an ℓ_∞-ball of radius D

$$\inf_{\text{algo}} \sup_{\text{functions } f} \varepsilon_n(\text{algo}, f) \geqslant \text{ cst } \times \min\left\{ \frac{B^2}{\mu n}, \frac{B^2}{\mu d}, BD\sqrt{\frac{d}{n}}, BD \right\}$$

  - The upper-bound is obtained through through stochastic subgradient: they match!

4.13.1 Sketch of proof

The proof follows the following argument:

1. **Create a subclass of functions** indexed by some vertices α^j, $j = 1, \dots, M$ of the hypercube $\{-1, 1\}^d$, which are sufficiently far in Hamming metric Δ_H (denote $\mathcal{V}$ this set with $|\mathcal{V}| = M$)

$$\forall j \neq k,\ \Delta_H(\alpha^i, \alpha^j) \geqslant \frac{d}{4},$$

   e.g., a "$\frac{d}{4}$-packing" (possible with M exponential in d - see later).

   The Hamming metric is defined through $\Delta_H(\alpha, \beta) = \sum_{i=1}^d 1_{\alpha_i \neq \beta_i}$.

2. **Design convex functions** so that

   – approximate optimization of the function is equivalent to function identification among the class above

   – stochastic oracle corresponds to a sequence of coin tosses with biases index by α^j, $j = 1, \dots, M$

   Thus the existence of an algorithm with a given convergence rate implies the existence of an algorithm identifying which of the M coins has been used for the coin tosses (the larger the M the harder such an algorithm is to be found).

3. Any such identification procedure (i.e., **a test**) has a lower bound on the probability of error.

4.13.2 Packing number for the hyper-cube

This is simply the **Varshamov-Gilbert's lemma** (Massart, 2003, p. 105): the maximal number of points in the hypercube that are at least $d/4$-apart in Hamming loss is greater than than $\exp(d/8)$.

The proof is as follows:

1. Maximality of family: if we take a maximal family, it has to satisfy $\mathcal{V} \Rightarrow \bigcup_{\alpha \in \mathcal{V}} \mathcal{B}_H(\alpha, d/4) = \{-1, 1\}^d$ otherwise we can find a new point which is at least $d/4$ apart from the existing ones.

2. Cardinality: using the union bound, we get $\sum_{\alpha \in \mathcal{V}} |\mathcal{B}_H(\alpha, d/4)| \geqslant 2^d$.

3. We can then link the cardinality of single set with the deviation of Z distributed as Binomial$(d, 1/2)$ (i.e., the sum of d unbiased Bernoulli random variables):

$$2^{-d}|\mathcal{B}_H(\alpha, d/4)| = \mathbb{P}(Z \leqslant d/4) = \mathbb{P}(Z \geqslant 3d/4)$$

4. Since every Bernoulli is between 0 and 1, we can apply Hoeffding's inequality:

$$\mathbb{P}(Z - \frac{d}{2} \geqslant \frac{d}{4}) \leqslant \exp(-\frac{2(d/4)^2}{d}) = \exp(-\frac{d}{8}).$$

   This implies that $1 \leqslant |\mathcal{V}| \exp(-d/8)$ and hence the result.

4.13.3 Designing a class of functions

Given $\alpha \in \{-1,1\}^d$, and a precision parameter $\delta > 0$, we consider the functions of the form:

$$g_\alpha(x) = \frac{c}{d}\sum_{i=1}^{d}\Big\{(\frac{1}{2}+\alpha_i\delta)f_i^+(x) + (\frac{1}{2}-\alpha_i\delta)f_i^-(x)\Big\},$$

where the 1-Lipschitz-continuous convex functions f_i^+'s and f_i^-'s, and the constant c are here ensure proper regularity and/or strong convexity (which we will do later). We consider the following oracle:

(a) Pick an index $i \in \{1,\dots,d\}$ at random

(b) Draw $b_i \in \{0,1\}$ from a Bernoulli with parameter $\frac{1}{2}+\alpha_i\delta$

(c) Consider $\hat{g}_\alpha(x) = c\big[b_i f_i^+ + (1-b_i)f_i^-\big]$ and its value and gradient. We have by design: $\mathbb{E}\hat{g}_\alpha = g_\alpha$, i.e., a stochastic gradient.

4.13.4 Optimizing is function identification

The goal is to make sure that if g_α is optimized up to error ε, then this identifies $\alpha \in \mathcal{V}$. This requires the definition of a certain "metric" between functions:

$$\rho(f,g) = \inf_{\theta\in\mathcal{C}} f(\theta) + g(\theta) - \inf_{\theta\in\mathcal{C}} f(\theta) - \inf_{\theta\in\mathcal{C}} g(\theta),$$

for which $\rho(f,g) \geqslant 0$ with equality iff f and g have the same minimizers.[1]

Moreover, we have the following lemma.

Lemma 4.1 *let $\psi(\delta) = \min_{\alpha\neq\beta\in\mathcal{V}} \rho(g_\alpha, g_\beta)$. For any $\tilde{\theta} \in \mathcal{C}$, there is at most one function g_α such that $g_\alpha(\tilde{\theta}) - \inf_{\theta\in\mathcal{C}} g_\alpha(\theta) \leqslant \frac{\psi(\delta)}{3}$.*

Proof Let $\tilde{\theta} \in \mathcal{C}$ such that $g_\alpha(\tilde{\theta}) - \inf_{\theta\in\mathcal{C}} g_\alpha(\theta) \leqslant \frac{\psi(\delta)}{3}$. By definition of $\psi(\delta)$, for any $\beta \neq \alpha$, we have:

$$\psi(\delta) \leqslant g_\alpha(\tilde{\theta}) + g_\beta(\tilde{\theta}) - \inf_{\theta\in\mathcal{C}} g_\alpha(\theta) - \inf_{\theta\in\mathcal{C}} g_\beta(\theta) \leqslant \frac{\psi(\delta)}{3} + g_\beta(\tilde{\theta}) - \inf_{\theta\in\mathcal{C}} g_\beta(\theta),$$

which implies that $g_\beta(\tilde{\theta}) - \inf_{\theta\in\mathcal{C}} g_\beta(\theta) \geqslant 2\frac{\psi(\delta)}{3}$ and hence the result. ■

Therefore:

- (a) optimizing an unknown function from the class up to precision $\frac{\psi(\delta)}{3}$ leads to identification of $\alpha \in \mathcal{V}$.

[1] Proof: (a) non-negativity is obvious, (b) if the minimizers are the same, then $\rho(f,g) = 0$ is obvious, (c) if $\rho(f,g) = 0$, then the sum of the infima is equal to the infimum of the sum if the two infima are attained simultaneously.

- (b) If the expected minimax error rate is greater than $\frac{\psi(\delta)}{9}$, there exists a function from the set of random gradient and function values such the probability of error is less than 1/3. Indeed, given an algorithm with expected approximation error (in function values) less than $\frac{\psi(\delta)}{9}$, applied to g_{α^*}, i.e., we have found after n steps a θ_n such that $\mathbb{E}g_{\alpha^*}(\theta_n) - \inf_{\theta\in\mathcal{C}} g_{\alpha^*}(\theta) \leqslant \frac{\psi(\delta)}{9}$.

  We are going to build an estimator $\hat{\alpha}$.

  If there is an α such that $g_\alpha(\theta_n) - \inf_{\theta\in\mathcal{C}} g_\alpha(\theta) \leqslant \frac{\psi(\delta)}{3}$, we take $\hat{\alpha}$ as this α (from the lemma above, there can be only one). If no such α exits, we take $\hat{\alpha}$ uniformly at random.

  Thus if $g_{\alpha^*}(\theta_n) - \inf_{\theta\in\mathcal{C}} g_{\alpha^*}(\theta) \leqslant$, then $\hat{\alpha} = \alpha^*$. This implies that (by Markov inequality)

$$\mathbb{P}(\hat{\alpha} \neq \alpha^*) \leqslant \frac{3}{\psi(\delta)}\left[\mathbb{E}g_{\alpha^*}(\theta_n) - \inf_{\theta\in\mathcal{C}} g_{\alpha^*}(\theta)\right] \leqslant \frac{1}{3}.$$

4.13.5 Lower bounds on coin tossing

- **Lemma**: For $\delta < 1/4$, given α^* uniformly at random in $\mathcal{V}$, if n outcomes of a random single coin (out of the d) are revealed, then any test will have a probability of error greater than

$$1 - \frac{16n\delta^2 + \log 2}{\frac{d}{2}\log(2/\sqrt{e})}$$

  - Proof based on Fano's inequality: If g is a function of Y, and X takes m values, then

$$\mathbb{P}(g(X) \neq Y) \geqslant \frac{H(X|Y) - 1}{\log m} = \frac{H(X)}{\log m} - \frac{I(X,Y) + 1}{\log m}$$

  - See Agarwal et. al for details.

4.13.6 Construction of f_i for convex functions

We consider the following functions $f_i^+(\theta) = |\theta(i) + \frac{1}{2}|$ and $f_i^-(\theta) = |\theta(i) - \frac{1}{2}|$, optimized on $[-1/2, 1/2]^d$. They depend on a single coordinate, and they are 1-Lipschitz-continuous with respect to the ℓ_2-norm. We thus have an i-th partial derivative bounded by c/d. With $c = B/2$, this implies that g_α is $B/(2\sqrt{d})$-Lipschitz, and thus B-Lipschitz.

Moreover, any call to the oracle corresponds to observing one of the d coins.

We can then compute the following lower bound on the discrepancy function (see Agarwal et. al for details):

- Fact 1: each g_α is minimized at $\theta_\alpha = -\alpha/2$ (NB: optimizing g_α indeed reveals the coin)
- Fact 2: the minimal value is $c/2 - c\delta$
- Fact 3: $\rho(g_\alpha, g_\beta) = \frac{2c\delta}{d}\Delta_H(\alpha,\beta) \geqslant \frac{c\delta}{2} = \psi(\delta)$ (by definition of $\psi(\delta)$)
- Set error/precision $\varepsilon = \frac{c\delta}{18}$ so that $\varepsilon < \psi(\delta)/9$

- Consequence: $\frac{1}{3} \geqslant 1 - \frac{16n\delta^2 + \log 2}{\frac{d}{2}\log(2/\sqrt{e})}$, that is, $\boxed{n \geqslant \text{ cst } \times \frac{L^2 d^2}{\varepsilon^2}}$

4.13.7 Construction of f_i for strongly-convex functions

- $f_i^{\pm}(\theta) = \frac{1}{2}\kappa|\theta(i) \pm \frac{1}{2}| + \frac{1-\kappa}{4}(\theta(i) \pm \frac{1}{2})^2$
  - Strongly convex and Lipschitz-continuous
- Same proof technique (more technical details)
- See more details by Agarwal et al. (2012); Raginsky and Rakhlin (2011)

Bibliography

[1] Lin Xiao. *Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization.* Journal of Machine Learning Research 11 (2010) 2543-2596.

[2] Amir Beck, Marc Teboulle. *Mirror descent and nonlinear projected subgradient methods for convex optimization.* Operations Research Letters 31 (2003) 167 – 175.

[3] Arkadi Nemirovski. *Tutorial: Mirror Descent Algorithms for Large-Scale Deterministic and Stochastic Convex Optimization.* Edinburgh, June 24-27, 2012.

[4] Sébastien Bubeck. *Convex Optimization: Algorithms and Complexity.* Foundations and Trends in Machine Learning, Vol. 8, No. 3-4 (2015) 231–358 2015

[5] Angelia Nedic and Soomin Lee. *On Stochastic Subgradient Mirror-Descent Algorithm with Weighted Averaging.* July 22, 2013.

[6] Harold Kushner, George Yin. *Stochastic Approximation and Recursive Algorithms and Applications (Stochastic Modelling and Applied Probability).*

[7] Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. *On the Generalization Ability of On-Line Learning Algorithms.*