10-725 Convex Optimization Course Notes

Instructor: Yuanzhi Li
Notes by Yan Pan
Carnegie Mellon University

Also known as Optimization for Machine Learning/Non-Convex Optimization.

1 Introduction

Definition 1.1 (Convex Set)**.** A set $\mathcal{D}$ is *convex* if for every $x, y \in \mathcal{D}$ and $0 \leq \lambda \leq 1$, we have

$$(1-\lambda)x + \lambda y \in \mathcal{D}.$$

Definition 1.2 (Convex Function)**.** A function f is *convex* over a convex set $\mathcal{D}$ if for every $x, y \in \mathcal{D}$ and $0 \leq \lambda \leq 1$, we have

$$f\left((1-\lambda)x + \lambda y\right) \leq (1-\lambda)f(x) + \lambda f(y).$$

Remark 1.2.1 (Alternative Definition of Convexity)**.** A second order differentiable function over a convex set $\mathcal{D}$ is *convex* iff for every vector v and every $x \in \mathcal{D}$, we have

$$v^\top \nabla^2 f(x) v \geq 0.$$

Lemma 1.3 (Lower Linear Bound)**.** *For every differentiable convex function f over a convex set $\mathcal{D}$ and every $x, y \in \mathcal{D}$, we have*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Proof. Proof by contradiction. Assume for contradiction that for any $\varepsilon > 0$ exists $y \in \mathcal{D}$ such that $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle - \varepsilon$. Then by convexity, for all $0 \leq \lambda \leq 1$, we have

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y) \leq f(x) + \lambda \langle \nabla f(x), y - x \rangle - \lambda\varepsilon,$$

$$f(x + \lambda(y-x)) \leq f(x) + \lambda \langle \nabla f(x), y - x \rangle - \lambda\varepsilon$$

which is (for non-zero λ)

$$\frac{f(x + \lambda(y-x)) - f(x)}{\lambda} \leq \langle \nabla f(x), y - x \rangle - \varepsilon.$$

By definition,

$$\lim_{\lambda \to 0} \frac{f(x + \lambda(y-x)) - f(x)}{\lambda} = \langle \nabla f(x), y - x \rangle.$$

So let $\lambda \to 0^+$ we have $\langle \nabla f(x), y - x \rangle \leq \langle \nabla f(x), y - x \rangle - \varepsilon$, which is a contradiction. □

Definition 1.4 (Convex Optimization)**.** An optimization problem is considered as *convex optimization* if f is a convex function and $\mathcal{D}$ is a convex set.

Remark 1.4.1. A convex function does not need to be differentiable, such as $f(x) = |x|$.

Remark 1.4.2. Examples of convex optimization problems:

- Linear regression: $\min_x \|y - Ax\|_2^2$.
- Ridge regression: $\min_x \|y - Ax\|_2^2 + \lambda\|x\|_2^2$.
- Logistic regression: $\min_x \sum_i -\log \frac{1}{1+e^{-y_i\langle A_i, x\rangle}}$.

Theorem 1.5. *For a first order differentiable convex function f, $\nabla f(x^*) = 0$ iff $f(x^*) = \min_x f(x)$.*

Proof. ($\Rightarrow$) By the lower linear bound, for every $x \in \mathcal{D}$, $f(x) \geq f(x^*)$ since $\nabla f(x^*) = 0$.
($\Leftarrow$) Next lecture. □

Remark 1.5.1. Convex function is considered easy because it has no local minima.

Corollary 1.5.2. *For general Lipschitz function, $\exists y \in \partial f(x^*), y = 0$ iff $f(x^*) = \min_x f(x)$.*

2 Gradient Descent and Mirror Descent

Definition 2.1 (Gradient Descent Algorithm)**.** Given a starting point x_0 and a learning rate η, for $t = 0, 1, 2, ...$, we compute x_{t+1} according to

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

Remark 2.1.1. The selection of the learning rate is important in gradient descent algorithm. Intuitively, we might want to choose large learning rates for "smooth" functions and small learning rates for "steep" functions. This gives us the motivation to define the *L-smoothness* of functions. With the definition of L-smoothness, we can now derive a lemma for gradient descent about how much the function decreases.

Definition 2.2 (L-smoothness)**.** A first order differentiable function (not necessarily convex) f over a set (not necessarily convex) $\mathcal{D}$ is called *L-smooth* for some $L > 0$ if for every $x, y \in \mathcal{D}$, we have

$$f(y) \leq f(x) + \langle\nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2.$$

Remark 2.2.1. The above formula is also referred to as the *upper quadratic bound*. Now for convex function f, we have both the lower linear bound and upper quadratic bound, giving us for every $x, y \in \mathcal{D}$,

$$f(x) + \langle\nabla f(x), y - x\rangle \leq f(y) \leq f(x) + \langle\nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2.$$

Corollary 2.2.2 (Smoothness Under Addition)**.** *If f_1 and f_2 are L-smooth, then $f(x) = f_1(x) + f_2(x)$ is $2L$-smooth.*

Proof. Trivial. □

Remark 2.2.3 (Alternative Definition of L-smoothness)**.** A second order differentiable function over a convex set $\mathcal{D}$ is *L-smooth* if for all $x \in \mathcal{D}$, we have

$$L \geq \|\nabla^2 f(x)\|_{sp}$$

where $\|\cdot\|_{sp}$ is the *spectral norm*, defined as the largest singular value of the matrix. Or equivalently for every unit vector v

$$v^\top \nabla^2 f(x) v \leq L.$$

The proof that the two definitions are equivalent is omitted because it's highly non-trivial.

Corollary 2.2.4. *Every third order differentiable function over a closed, bounded convex set $\mathcal{D}$ is L-smooth for some finite L.*

Lemma 2.3 (Gradient Descent Lemma)**.** *For any L-smooth function f, as long as $\eta \leq \frac{1}{L}$, we have*

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2.$$

Proof. Notice that $x_{t+1} = x_t - \eta \nabla f(x_t)$ in gradient descent. By definition of L-smoothness, we have

$$f(x_{t+1}) \leq f(x_t) - \langle \nabla f(x_t), \eta \nabla f(x_t) \rangle + \eta^2 \frac{L}{2}\|\nabla f(x_t)\|_2^2.$$

Since $\eta^2 \frac{L}{2} \leq \eta$ for $\eta \leq \frac{1}{L}$, we have

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2$$

as desired. □

Lemma 2.4 (Convergence Rate of Gradient)**.** *In gradient descent, for every $\varepsilon > 0$, let*

$$T_\varepsilon = \frac{2(f(x_0) - \min_x f(x))}{\eta \varepsilon},$$

then there exists $t \leq T_\varepsilon$ such that $\|\nabla f(x_t)\|_2^2 \leq \varepsilon$.

Proof. By contradiction. Suppose for every $t \leq T_\varepsilon$, $\|\nabla f(x_t)\|_2^2 > \varepsilon$. Then,

$$f(x_{t+1}) < f(x_t) - \frac{\eta}{2}\varepsilon$$

which implies that

$$f(x_{T_\varepsilon}) < f(x_0) - \frac{\eta}{2} T_\varepsilon \varepsilon = f(x_0) - (f(x_0) - \min_x f(x)) = \min_x f(x)$$

which is a contradiction. □

Remark 2.4.1. Lemma 2.4 shows that the learning rate does not depend on how large the gradient is, although in practice we do want to tune the learning rate for faster convergence.

Remark 2.4.2. Lemma 2.4 shows that we can decrease the objective, however if the gradient is too small, it might take forever for gradient to reach exactly zero. Take the function $f(x) = \varepsilon^2 x^2$ as an example, when $x = \frac{1}{\varepsilon}$ we have $|\nabla f(x)| = \varepsilon$ but $f(x) = 1$. Therefore, we also need to study the convergence rate of gradient descent.

Lemma 2.5 (Mirror Descent Lemma)**.** *In gradient descent, for any L-smooth function f and $y \in \mathcal{D}$, we have*

$$f(x_t) \leq f(y) + \frac{1}{2\eta}\left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2 + \|x_{t+1} - x_t\|_2^2\right).$$

Proof. By lower linear bound we have

$$f(y) \geq f(x_t) + \langle \nabla f(x_t), y - x_t \rangle = f(x_t) + \frac{1}{\eta}\langle x_t - x_{t+1}, y - x_t \rangle.$$

We observe that

$$\begin{aligned}\langle x_t - x_{t+1}, y - x_t \rangle &= \langle x_{t+1}, x_t \rangle + \langle y, x_t \rangle - \langle x_{t+1}, y \rangle - \langle x_t, x_t \rangle \\ &= -\frac{1}{2}\left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2 + \|x_{t+1} - x_t\|_2^2\right).\end{aligned}$$

Hence we have the desired result. □

Remark 2.5.1. The mirror descent lemma implies that if the current function value $f(x_t)$ is much larger than $f(x^*)$, then since $\|x_{t+1} - x_t\|_2^2$ is small, the values $\|x^* - x_{t+1}\|_2^2$ must be much smaller than $\|x^* - x_t\|_2^2$, so x_{t+1} will be much closer to x^* compared to x_t. So the mirror descent lemma looks at decreasing the distance between x_t to x^*.

Theorem 2.6 (Convergence Rate of Gradient Descent)**.** *In gradient descent, as long as $\eta \leq \frac{1}{L}$, we have*

$$f(x_T) \leq f(x^*) + \frac{\|x^* - x_0\|_2^2}{\eta T}.$$

Proof. Let $y = x^*$, summing up the mirror descent lemma for $t = 0, \cdots, t-1$, we have the *telescoping sum* inequality

$$\sum_{t=0}^{T-1} f(x_t) \leq T f(x^*) + \frac{1}{2\eta}\left(\|x^* - x_0\|_2^2 - \|x^* - x_T\|_2^2 + \sum_{t=0}^{T-1} \|x_{t+1} - x_t\|_2^2\right)$$

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \leq f(x^*) + \frac{1}{2\eta T}\left(\|x^* - x_0\|_2^2 - \|x^* - x_T\|_2^2 + \sum_{t=0}^{T-1} \|x_{t+1} - x_t\|_2^2\right)$$

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \leq f(x^*) + \frac{1}{2\eta T}\left(\|x^* - x_0\|_2^2 + \sum_{t=0}^{T-1} \|x_{t+1} - x_t\|_2^2\right)$$

which implies using $x_{t+1} = x_t - \eta \nabla f(x_t)$

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \leq f(x^*) + \frac{1}{2\eta T}\|x^* - x_0\|_2^2 + \frac{\eta}{2T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2.$$

When $\eta \leq \frac{1}{L}$, by Gradient Descent Lemma we have

$$\frac{\eta}{2}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2 \leq f(x_0) - f(x_T) \leq f(x_0) - f(x^*)$$

since x^* is a minimizer of f. Then,

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \leq f(x^*) + \frac{1}{T}\left(\frac{1}{2\eta}\|x^* - x_0\|_2^2 + f(x_0) - f(x^*)\right).$$

Using the L-smoothness of f that shows $f(x_0) - f(x^*) \leq \frac{L}{2}\|x^* - x_0\|_2^2$ and the gradient descent lemma that shows $f(x_T) \leq f(x_t)$, we have

$$f(x_T) \leq f(x^*) + \frac{\|x^* - x_0\|_2^2}{T}\left(\frac{1}{2\eta} + \frac{L}{2}\right)$$

which implies when $\eta \leq \frac{1}{L}$

$$f(x_T) \leq f(x^*) + \frac{\|x^* - x_0\|_2^2}{\eta T}.$$

□

3 Momentum

The intuition behind *accelerated gradient descent* is to use larger learning rate $\eta > \frac{1}{L}$ without entering zig-zag. Some functions are only non-smooth at some corners, and using a low learning rate $\frac{1}{L}$ might not be the most efficient option.

The key idea of momentum is to use a universal large learning rate and use the "weighted" sum of the gradients from the previous iterations to update the current point. When gradients point to the same direction, the sum will be large. Otherwise, when the gradients bump back and forth, the sum will be small. The weighted sum of the past gradients is called the *momentum*.

Definition 3.1 (Momentum)**.** Use a learning rate $\eta > \frac{1}{L}$, and update using

$$x_{t+1} = x_t - \eta g_t.$$

Definition 3.2 (Nesterov's Accelerated Gradient Descent)**.** For a L-smooth function f, in each iteration, we compute

$$z_{t+1} = x_t - \eta \nabla f(x_t)$$

$$x_{t+1} = (1-\gamma_t) z_{t+1} + \gamma_t z_t$$

$$\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}, \qquad \gamma_t = \frac{1-\lambda_t}{\lambda_{t+1}}.$$

Remark 3.2.1. Nesterov's accelerated gradient descent is guaranteed to work mathematically, but in practice it's not the best momentum. In contrast, people use the *heavy ball momentum* more often, for example in PyTorch `optim.SGD(momentum=0.9)`, but it doesn't have theoretical guarantee.

Definition 3.3 (Heavy Ball Momentum)**.** The update can be approximated by

$$x_{t+1} \approx x_t - \eta g_t$$

$$g_t = \gamma \sum_{s \le t} (1-\gamma)^{t-s} \nabla f(x_s)$$

where the last step can be updated easily using

$$g_{t+1} = g_t(1-\gamma) + \gamma \nabla f(x_{t+1}).$$

Therefore, we can choose $\eta = \frac{1}{\gamma L} > \frac{1}{L}$.

Remark 3.3.1. There is no proof for why heavy ball momentum works, but we can do a thought experiment to show why this works intuitively. The key observation is that when gradient is smaller than usual, we can use a larger learning rate. We fix a value $K > 0$, if $\|f(x_t)\|_2^2 \ge K$ holds for *every* t, using $\eta = \frac{1}{L}$ and the gradient descent lemma, we have

$$f(x_{t+1}) \le f(x_t) - \frac{K}{2L}$$

so we need at most $\frac{Lf(x_0)}{K}$ iterations to find a point x_T with $f(x_T) \le \frac{f(x_0)}{2}$. If $\|f(x_t)\|_2^2 < K$ holds for *every* t, using the telescoping sum of mirror descent lemma, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) \le \frac{1}{2\eta T}\|x^* - x_0\|_2^2 + \frac{\eta K}{2}$$

with the assumption that $f(x^*) = 0$. With $\eta = \frac{f(x_0)}{2K}$, we need at most $\frac{4K\|x_0 - x^*\|_2^2}{f(x_0)^2}$ iterations to find a point x_T with $f(x_T) \leq \frac{f(x_0)}{2}$.

Therefore, picking $K = \sqrt{\frac{Lf^3(x_0)}{4\|x_0 - x^*\|_2^2}}$, in both cases we need at most $\frac{2\|x_0 - x^*\|_2\sqrt{L}}{\sqrt{f(x_0)}}$ iterations to find a point x_T with $f(x_T) \leq \frac{f(x_0)}{2}$. In the second case, when $f(x_0) \approx \|x_0 - x^*\|_2 \approx 1$, the learning rate is indeed much larger $\eta = \frac{f(x_0)}{K} \approx \frac{1}{\sqrt{L}} > \frac{1}{L}$. Therefore, the convergence of the thought experiment is, for $\varepsilon > 0$, we need at most $\frac{\sqrt{2L}}{\sqrt{\varepsilon}}$ iterations to find a point x_{T_ε} with $f(T_\varepsilon) \leq \varepsilon$.

Definition 3.4 (Linear Coupling)**.** For a $0 \leq \tau \leq 1$, at every iteration, we compute

$$s_{t+1} = x_t - \frac{1}{L}\nabla f(x_t) \qquad \text{(gradient descent)}$$

$$l_{t+1} = l_t - \eta \nabla f(x_t). \qquad \text{(momentum)}$$

The updated value is

$$x_{t+1} = (1 - \tau)s_{t+1} + \tau l_{t+1}.$$

Remark 3.4.1. Linear coupling is a combination of gradient descent and momentum. We need this because in the thought experiment, it might be the case that neither $\|\nabla f(x_t)\|_2^2 \geq K$ and $\|\nabla f(x_t)\|_2^2 < K$ holds for *every* t. Linear coupling gives us a better t.

4 Constraint Optimization

Definition 4.1 (Constraint Convex Optimization)**.** A convex optimization function is a *constraint convex optimization* if the set $\mathcal{D}$ is a strict subset of $\mathbb{R}^d$.

Remark 4.1.1. A problem with gradient descent in the constraint convex optimization problem is $x_t \in \mathcal{D}$ does not necessarily imply $x_{t+1} \in \mathcal{D}$.

Definition 4.2 (Algorithmic Projected Gradient Descent)**.** The update rule for convex set $\mathcal{D}$ is

$$x_{t+1} = \Pi_{\mathcal{D}}(x_t - \eta \nabla f(x_t)),$$

where

$$\Pi_{\mathcal{D}}(x) = \arg\min_z \|z - x\|_2^2.$$

Remark 4.2.1. For many convex sets $\mathcal{D}$, the projection is not easy to compute. For example, the polytope. There will be methods such as the *min-max optimization algorithm* and the *interior point method* to solve them.

Lemma 4.3. *For a convex set $\mathcal{D}$, for every $x \in \mathbb{R}^d$, the projection $\Pi_{\mathcal{D}}(x)$ is unique.*

Proof. If $x \in \mathcal{D}$, $\arg\min_z \|z - x\|_2^2 = x$. If $x \notin \mathcal{D}$, suppose $z_1, z_2 \in \mathcal{D}$, $z_1 \neq z_2$ such that $\|z_1 - x\|_2^2 = \|z_2 - x\|_2^2$, then $z' = \frac{z_1 + z_2}{2} \in \mathcal{D}$ by convexity of $\mathcal{D}$. By convexity of $f(z) = \|z - x\|_2^2$,

$$\|z' - x\|_2^2 < \frac{\|z_1 - x\|_2^2 + \|z_2 - x\|_2^2}{2}.$$

Therefore z_1 and z_2 are not the projection. □

Definition 4.4 (Gradient Mapping)**.** In the algorithmic projected gradient descent, the *gradient mapping* is defined as

$$g(x_t) = \frac{1}{\eta}\left(x_t - \Pi_{\mathcal{D}}(x_t - \eta \nabla f(x_t))\right)$$

and the update rule becomes

$$x_{t+1} = x_t - \eta g(x_t).$$

Lemma 4.5 (Lower Linear Bound for Convex Set)**.** *For every $y' \notin \mathcal{D}$ and $y \in \mathcal{D}$,*

$$\langle y' - \Pi_{\mathcal{D}}(y'), \Pi_{\mathcal{D}}(y') - y \rangle \geq 0.$$

Lemma 4.6 (Gradient Descent Lemma for Gradient Mapping)**.** *For convex set $\mathcal{D}$ and L-smooth function f, when $\eta \leq \frac{1}{L}$,*

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\langle \nabla f(x_t), g(x_t) \rangle$$

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|g(x_t)\|_2^2.$$

Proof. We apply the lower linear bound for convex set

$$\langle x_t - \eta \nabla f(x_t) - x_{t+1}, x_{t+1} - x_t \rangle \geq 0$$

$$\langle -\eta \nabla f(x_t) + \eta g(x_t), -\eta g(x_t) \rangle \geq 0$$

which implies that

$$\langle \nabla f(x_t), g(x_t) \rangle \geq \|g(x_t)\|_2^2.$$

When the function is L-smooth, we have

$$f(x_{t+1}) \leq f(x_t) - \eta \langle \nabla f(x_t), g(x_t) \rangle + \frac{L\eta^2}{2}\|g(x_t)\|_2^2.$$

Therefore, using $\langle \nabla f(x_t), g(x_t) \rangle \geq \|g(x_t)\|_2^2$ we derive the desired result. □

Lemma 4.7 (Mirror Descent Lemma for Gradient Mapping)**.** *For any $y \in \mathcal{D}$, we have*

$$f(x_t) \leq f(y) + \frac{1}{2\eta}\left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2 + 2\eta^2 \langle \nabla f(x_t), g(x_t) \rangle\right).$$

5 Mirror Descent

Definition 5.1 (Bregman Divergence)**.** For a differentiable convex function g, the *Bregman divergence* is defined as

$$D_g(x, y) = g(x) - g(y) - \langle \nabla g(y), x - y \rangle.$$

Remark 5.1.1. Bregman divergence is a large class of distances. When $g(x) = \|x\|_2^2$, then $D_g(x, y) = \|x - y\|_2^2$, which is the *Euclidean distance*. When $g(x) = \sum_{i \in [d]} x_i \log x_i$ where each $x_i \geq 0$ and $\sum_{i \in [d]} x_i = \sum_{i \in [d]} y_i = 1$,

$$\begin{aligned} D_g(x, y) &= g(x) - g(y) - \langle \nabla g(y), x - y \rangle \\ &= \sum_{i \in [d]} (x_i \log x_i - y_i \log y_i) - \sum_{i \in [d]} (\log y_i + 1)(x_i - y_i) \\ &= \sum_{i \in [d]} \left(x_i \log \frac{x_i}{y_i} + (x_i - y_i) \right) \\ &= \sum_{i \in [d]} x_i \log \frac{x_i}{y_i} \end{aligned}$$

which is the *KL-divergence*.

Definition 5.2 (Mirror Descent Algorithm)**.** Given a distance $D_g(x, y)$ defined by a differentiable convex function g, the *mirror descent algorithm* to minimize a function f is given by

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

Remark 5.2.1. As long as $g(x) = \omega(\|x\|_2)$ when $\|x\|_2 \to \infty$, such x_{t+1} can always be found. The criteria is sufficient but not necessary.

Lemma 5.3 (Real Mirror Descent Lemma)**.** *For a L-Lipschitz function f*

$$f(x_t) \leq f(y) + \frac{1}{\eta} \left(D_g(x^*, x_t) - D_g(x^*, x_{t+1}) + D_g(x_t, x_{t+1}) \right)$$

Definition 5.4 (L-Lipschitzness)**.** A function f is called L-Lipschitz with respect to a distance D_g if for every x, y, we have

$$|\langle \nabla f(x), y - x \rangle| \leq L \sqrt{D_g(y, x)}.$$

Specifically, for $D_g(y, x) = \|y - x\|_2^2$, we have

$$\|\nabla f(x)\|_2 \leq L$$

or equivalently,

$$|f(y) - f(x)| \leq L \|y - x\|_2.$$

6 Stochastic Gradient Descent

Definition 6.1 (Empirical Risk Minimization)**.** Given training data set $\{x_i, y_i\}_{i=1}^N$ where x_i's are the *training data*, y_i's are the *training labels*, the *ERM* type of problem is given as

$$\min_W \frac{1}{N} \sum_{i=1}^N \ell(h(x_i, W), y_i) + R(W)$$

where h is a *parameterized model*, W is the *trainable parameters*, ℓ is the *loss function*, and R is the *regularizer*.

Definition 6.2 (Stochastic Gradient Descent)**.** In the same setting as gradient descent, the *stochastic gradient descent* minimizes a function as

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

where $\mathbb{E}\left[\tilde{\nabla} f(x_t)\right] = \nabla f(x_t)$.

Lemma 6.3 (Expectation of Stochastic Gradient)**.** *If we sample a subset $\mathcal{S}_t$ uniformly at random, the expectation of the stochastic gradient is the true gradient.*

Proof. Observe that $\mathbb{E}\left[\mathbf{1}_{i \in \mathcal{S}_t}\right] = \mathbb{P}(i \in \mathcal{S}_t) = \frac{m}{N}$ for any i since the subset is sampled uniformly at

random. Then,

$$\begin{aligned}
\mathbb{E}\left[\tilde{\nabla} f(W_t)\right] &= \mathbb{E}\left[\frac{1}{m}\sum_{i\in\mathcal{S}_t}\nabla\ell(h(x_i,W_t),y_i)+\nabla R(W_t)\right] \\
&= \frac{1}{m}\mathbb{E}\left[\sum_{i\in\mathcal{S}_t}\nabla\ell(h(x_i,W_t),y_i)\right]+\nabla R(W_t) \\
&= \frac{1}{m}\mathbb{E}\left[\sum_{i\in[N]}\mathbf{1}_{i\in\mathcal{S}_t}\nabla\ell(h(x_i,W_t),y_i)\right]+\nabla R(W_t) \\
&= \frac{1}{m}\sum_{i\in[N]}\mathbb{E}\left[\mathbf{1}_{i\in\mathcal{S}_t}\right]\nabla\ell(h(x_i,W_t),y_i)+\nabla R(W_t) \\
&= \frac{1}{m}\sum_{i\in[N]}\frac{m}{N}\nabla\ell(h(x_i,W_t),y_i)+\nabla R(W_t) \\
&= \frac{1}{N}\sum_{i\in[N]}\nabla\ell(h(x_i,W_t),y_i)+\nabla R(W_t) \\
&= \nabla f(W_t).
\end{aligned}$$

□

7 Duality and Min-Max Optimization

Definition 7.1 (Lagrange Duality)**.** For some differentiable functions $h_1,\cdots,h_m,l_1,\cdots,l_n$, given the constraint set

$$\mathcal{D}=\{x\in\mathbb{R}^d \mid \forall i\in[m], h_i(x)\le 0; \forall j\in[n], l_j(x)=0\}.$$

Let

$$L(x,u,v)=f(x)+\sum_{i\in[m]}u_ih_i(x)+\sum_{j\in[n]}v_jl_j(x),$$

then the *Lagrange dual function* is defined as

$$g(u,v)=\min_{x\in\mathbb{R}^d}L(x,u,v)$$

Theorem 7.2. *To solve* $\min f(x)$*,* $x\in\mathcal{D}$ *where* $\mathcal{D}$ *defined the same as above, one can alternatively solve*

$$\min_{x\in\mathbb{R}^d}\max_{u\ge 0,v}L(x,u,v).$$

Theorem 7.3 (Karush-Kuhn-Tucker Condition)**.** *Let* x^*,u^*,v^* *be a solution of*

$$\max_{u\ge 0,v}\min_{x\in\mathbb{R}^d}L(x,u,v)$$

then the following conditions must hold

$$\nabla_x L(x^*,u^*,v^*)=\nabla f(x^*)+\sum_{i\in[m]}u_i^*\nabla h_i(x^*)+\sum_{j\in[n]}v_j^*\nabla l_j(x^*)=0 \qquad \text{(stationarity)}$$

$$\forall i\in[m], u_i^*h_i(x^*)=0 \qquad \text{(complementary slackness)}$$

$$x^*\in\mathcal{D} \qquad \text{(primal feasibility)}$$

$$u^*\ge 0. \qquad \text{(dual feasibility)}$$