DSC 211 Introduction to Optimization Winter 2024 Instructor: Jun-Kun Wang
Scribe: Merlin Chang, Marialena Sfyraki February 13, 2024

Lecture 11: Fenchel Conjugate, Dual formulation of the Empirical Risk Minimization, and SDCA

1 Fenchel Conjugate

Definition 1 (Fenchel Conjugate)**.** *Consider a function $f(\cdot)$, then the Fenchel Conjugate is defined to be*

$$f^*(y) = \sup_{x \in \mathrm{dom}(f)} \left(y^\top x - f(x) \right).$$

Claim. *The conjugate function $f^*(y)$ is always convex, even if $f(\cdot)$ is non-convex.*

Proof. Let $h_x(y) := y^\top x - f(x)$. Observe that h_x is an affine function of y and therefore also convex. Let $\alpha \in [0,1]$ and $y_1, y_2 \in \mathrm{dom}(f^*)$. Then, we have

$$\begin{aligned} f^*\left((1-\alpha)y_1 + \alpha y_2\right) &= \sup_{x \in \mathrm{dom}(f)} h_x\left((1-\alpha)y_1 + \alpha y_2\right) \\ &= \sup_{x \in \mathrm{dom}(f)} (1-\alpha)h_x(y_1) + \alpha h_x(y_2) \\ &\le (1-\alpha) \sup_{x \in \mathrm{dom}(f)} h_x(y_1) + \alpha \sup_{x \in \mathrm{dom}(f)} h_x(y_2) \\ &= (1-\alpha) f^*(y_1) + \alpha f^*(y_2). \end{aligned}$$

Thus, by the zero-order characterization of convexity, we have that f^* is convex. □

Exercise 1. $f(x) = a^\top x + b$.

$$\begin{aligned} f^*(y) &= \sup_{x \in \mathrm{dom}(f)} \left(\langle y, x \rangle - f(x) \right) \\ &= \sup_{x \in \mathrm{dom}(f)} \left(\langle y - a, x \rangle - b \right) \\ &= \begin{cases} -b & \text{, if } y = a \\ \infty & \text{, otherwise} \end{cases}. \end{aligned}$$

Exercise 2. $f(x) = \frac{1}{2}x^2$.

$$
\begin{aligned}
f^*(y) &= \sup_{x \in \text{dom}(f)} \big(\langle y, x\rangle - f(x)\big) \\
&= \sup_{x \in \text{dom}(f)} \left(\langle y, x\rangle - \frac{1}{2}x^2\right)
\end{aligned}
$$

Let $h(x) := \langle y, x\rangle - \frac{1}{2}x^2$. Then, the maximizer of h can be found as

$$\nabla h(x) = 0 \;\Leftrightarrow\; x = y$$

Thus,

$$f^*(y) = x^2 - \frac{1}{2}x^2 = \frac{1}{2}x^2$$

1.1 Fenchel inequality

By the definition of the conjugate function, we have the following result:

Theorem 1 (Fenchel inequality)**.** *For any x and y, we have*

$$f^*(y) \geq y^\top x - f(x).$$

Question. When do we have the equality?

$$f^*(y) + f(x) = y^\top x$$

In the following, we are going to answer this equation and prove the following theorem:

Theorem 2. *If $f(\cdot)$ is closed and convex, then the following are equivalent:*

i. $f^*(y) + f(x) = y^\top x$.

ii. $y = \nabla f(x)$.

iii. $x = \nabla f^*(y)$.

Now, recall the definition of open and closed sets.

Definition 2 (Open set)**.** *A set S is open if it contains an open ball about each of its points. That is, for all $x \in S$, there exists $\epsilon > 0$ such that $B_\epsilon(x) \subset S$.*

Definition 3 (Closed set)**.** *A set S is closed if its complement is open.*

We will now introduce the definition of closed functions.

Definition 4 (Closed function)**.** *A function is closed if its sublevel set is a closed set, i.e.,*

$$\{x \in dom(f) : f(x) \leq \alpha\}$$

is a closed set.

Counterexample. $f(x) = \exp(-x)$ is not a closed function. Observe that its sublevel set $\{x \in \mathrm{dom}(f) : \exp(-x) \leq \alpha\}$ is not closed.

1.2 The inverse of the gradient map

Theorem 3. *Suppose that $f(\cdot)$ is closed and convex. Then, $y \in \partial f(x)$ if and only if $x \in \partial f^*(y)$.*

Proof. We will only prove the "$\Rightarrow$" direction, that is we will show that if $y \in \partial f(x)$ then $x \in \partial f^*(y)$. Let $y \in \partial f(x)$. By the first-order characterization of convexity, for any $u \in \mathbb{R}^d$ we have

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

Additionally, we have

$$\begin{aligned} f^*(y) &= \sup_u \big(\langle u, y\rangle - f(u)\big) && \text{(by definition of conjugate function)} \quad (1)\\ &\leq \sup_u \langle u, y\rangle - \big(f(x) + \langle y, u - x\rangle\big) && \text{(by convexity)} \quad (2)\\ &= \langle x, y\rangle - f(x) && (3) \end{aligned}$$

Recall that for a convex function $h(\cdot)$ defined over a convex set C, a vector g_x is said to be a sub-gradient of $f(\cdot)$ at a point $x \in C$ if for any $y \in C$

$$h(y) \geq h(x) + \langle g_x, y - x\rangle.$$

Now, for any $z \in \mathbb{R}^d$ we have

$$\begin{aligned} f^*(z) &\geq \langle z, x\rangle - f(x) && \text{(by definition of the Fenchel inequality)}\\ &= \langle z - y, x\rangle - f(x) + \langle y, x\rangle \\ &\geq \langle z - y, x\rangle + f^*(y) && \text{(by inequality (3)} \end{aligned}$$

By the fact that $f^*(\cdot)$ is convex (and differentiable) and by the definition of the subgradient we have that

$$x \in \partial f^*(y),$$

which concludes the proof.

To prove the other direction, we follow the same lines of the proof as above. Specifically, we let $r := f^*$ and the function $r(\cdot)$ is convex and closed. So we can use the above argument for $r(\cdot)$ and deduce that if $x \in \partial r(y)$, then $y \in \partial r^*(x)$. Then, using the fact that if $f(\cdot)$ is closed and convex, then the bi-conjugate $f^{**}(x)$ is equal to the original function $f(\cdot)$ itself, we can complete the proof

□

Question 1. What is $\arg\sup_{x \in \mathrm{dom}(f)} \big(y^\top x - f(x)\big)$ when $f(\cdot)$ is closed and convex ? This is because the $\arg\sup_{x \in \mathrm{dom}(f)} \big(y^\top x - f(x)\big)$ is what makes the Fenchel inequality becomes the equality.

Theorem 4. *Let $f(\cdot)$ be convex. We have*

$$f^*(y) + f(x) = y^\top x \iff y \in \partial f(x).$$

Proof. Let us first show that $f^*(y) + f(x) = y^\top x \Rightarrow y \in \partial f(x)$.

$$\begin{aligned} f^*(y) &= \sup_{x \in \mathrm{dom}(f)} \langle y, x\rangle - f(x) && (4) \\ &\geq \langle y, z\rangle - f(z), \quad \forall z \in \mathrm{dom}(f) && (5) \end{aligned}$$

Also, from $f^*(y) + f(x) = y^\top x$, we have

$$\begin{aligned} 0 &= f^*(y) + f(x) - y^\top x && (6) \\ &\overset{(5)}{\geq} \langle y, z\rangle - f(z) + f(x) - \langle y, x\rangle, \quad \forall z \in \mathrm{dom}(f) && (7) \end{aligned}$$

Hence, rearranging the above terms, we get

$$f(z) \geq f(x) + \langle y, z - x\rangle, \quad \forall z \in \mathrm{dom}(f), \tag{8}$$

which by the definition of the subgradient, we can conlude the $y \in \partial f(x)$.

Now let us prove the other direction $y \in \partial f(x) \Rightarrow f^*(y) + f(x) = y^\top x$.

From $y \in \partial f(x)$, we have

$$\begin{aligned} & f(z) \geq f(x) + \langle y, z - x\rangle, \forall z, x \in \mathrm{dom}(f) \\ \Rightarrow\ & \langle y, x\rangle - f(x) \geq \langle y, z\rangle - f(z), \forall z, x \in \mathrm{dom}(f) \\ \Rightarrow\ & \langle y, x\rangle - f(x) \geq \sup_{z \in \mathrm{dom}(f)} \langle y, z\rangle - f(z), \quad \forall x \in \mathrm{dom}(f) \\ & = f^*(y) && (9) \end{aligned}$$

On the other hand, by the definition of the conjugate function

$$f^*(y) \geq \langle y, x \rangle - f(x), \forall x \in \mathrm{dom}(f). \tag{10}$$

By combining the above, we can conclude that

$$f^*(y) + f(x) = y^\top x. \tag{11}$$

□

Now by combining Theorem 3 and Theorem 4, we know

$$\arg \sup_{x \in \mathrm{dom}(f)} \Big(y^\top x - f(x) \Big) \in \partial f^*(y). \tag{12}$$

and the following theorem:

Theorem 5. *If $f(\cdot)$ is closed and convex, then the following are equivalent:*

$$f^*(y) + f(x) = y^\top x \iff y \in \partial f(x) \iff x \in \partial f^*(y).$$

Question 2. What is $\arg\sup_{y \in \mathrm{dom}(f^*)} \big(y^\top x - f^*(y) \big)$ when $f(\cdot)$ is closed and convex?

Using a similar argument as Theorem 4, we can follow the same lines of its proof with one modification. Specifically, we let $r := f^*$ and the function $r(\cdot)$ is convex and closed. So we can use the above argument for $r(\cdot)$ and deduce that

$$\arg \sup_{x \in \mathrm{dom}(r)} \Big(y^\top x - r(x) \Big) \in \partial r^*(y). \tag{13}$$

Now using the fact that if $f(\cdot)$ is closed and convex, then the bi-conjugate $f^{**}(x)$ is equal to the original function $f(\cdot)$ itself, (13) leads to

$$\arg \sup_{x \in \mathrm{dom}(f^*)} \Big(y^\top x - f^*(x) \Big) \in \partial f(y). \tag{14}$$

2 Regularized Empirical Risk Minimization

If the primal problem is

$$\min_{x \in \mathbb{R}^d} F(x), \qquad \text{where } F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x^\top z_i) + \frac{\lambda}{2} \|x\|_2^2,$$

then the dual problem is

$$\max_{\alpha \in \mathbb{R}^n} D(\alpha), \qquad \text{where } D(\alpha) := \frac{1}{n} \sum_{i=1}^{n} -f_i^*(-\alpha_i) - \frac{\lambda}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i z_i \right\|_2^2.$$

We will show how the dual problem is derived from the primal problem. Consider the following constrained optimization problem

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} f_i(\theta_i) + \frac{\lambda n}{2}\|x\|_2^2$$
$$\text{subject to } \forall i, \theta_i = z_i^\top x,$$

where we have introduced variables $\{\theta_i\}_{i=1}^n$.

Step 1. Constructing the Lagrangian
The Lagrangian is formulated as

$$L\left(x, \{\theta_i\}, \{\alpha_i\}\right) = \sum_{i=1}^{n} \left[f_i(\theta_i) + \alpha_i\left(\theta_i - z_i^\top x\right)\right] + \frac{\lambda n}{2}\|x\|_2^2$$

Step 2. Optimizing over primal variables to get the dual function
We have that

$$\min_{x,\theta_1-\theta_n} \sum_{i=1}^{n} \left(f_i(\theta_i) + \alpha_i\theta_i - \alpha_i z_i^\top x\right) + \frac{\lambda n}{2}\|x\|_2^2$$
$$\iff \min_x \sum_{i=1}^{n} \left(\min_{\theta_i} f_i(\theta_i) + \alpha_i\theta_i\right) + \frac{\lambda n}{2}\|x\|_2^2 - \sum_{i=1}^{n} \alpha_i z_i^\top x.$$

Now, observe that

$$\min_\theta q(\theta) = -\max_\theta\left(-q(\theta)\right).$$

Thus, we have that

$$\begin{aligned}\left(\min_{\theta_i} f_i(\theta_i) + \alpha_i\theta_i\right) &= -\max_{\theta_i}\left[-\left(f_i(\theta_i) + \alpha_i\theta_i\right)\right]\\ &= -\max_{\theta_i}\left[-\alpha_i\theta_i - f_i(\theta_i)\right]\\ &= -f_i^*(\alpha_i) \qquad \text{(by definition of the conjugate)}\end{aligned}$$

Therefore, using the above result we can rewrite

$$\min_{x,\theta_1-\theta_n} \sum_{i=1}^{n} \left(f_i(\theta_i) + \alpha_i\theta_i - \alpha_i z_i^\top x\right) + \frac{\lambda n}{2}\|x\|_2^2$$
$$\iff \min_x \sum_{i=1}^{n} \left(\min_{\theta_i} f_i(\theta_i) + \alpha_i\theta_i\right) + \frac{\lambda n}{2}\|x\|_2^2 - \sum_{i=1}^{n} \alpha_i z_i^\top x$$
$$\iff -\sum_{i=1}^{n} f_i^*(-\alpha_i) + \min_x \underbrace{\frac{\lambda n}{2}\|x\|_2^2 - \sum_{i=1}^{n} \alpha_i z_i^\top x}_{q(x)}.$$

Additionally, observe that

$$q(x) = 0 \;\Leftrightarrow\; \lambda n x = \sum_{i=1}^{n} \alpha_i z_i \;\Leftrightarrow\; x = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i z_i$$

The equation

$$x = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i z_i$$

describes the **relation between primal variables and dual variables**. Using this result we have

$$\begin{aligned}
\min_x \frac{\lambda n}{2}\|x\|_2^2 - \sum_{i=1}^{n} \alpha_i z_i^\top x &= \frac{\lambda n}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i z_i \right\|_2^2 - \langle \sum_{i=1}^{n} \alpha_i z_i, \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i z_i \rangle \\
&= \frac{1}{2\lambda n} \left\| \sum_{i=1}^{n} \alpha_i z_i \right\|_2^2 - \frac{1}{\lambda n} \left\| \sum_{i=1}^{n} \alpha_i z_i \right\|_2^2 \\
&= -\frac{1}{2\lambda n} \left\| \sum_{i=1}^{n} \alpha_i z_i \right\|_2^2 \\
&= -\frac{\lambda n}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i z_i \right\|_2^2 .
\end{aligned}$$

Plugging this in the objective we get

$$\begin{aligned}
&\min_{x, \theta_1 - \theta_n} \sum_{i=1}^{n} \left(f_i(\theta_i) + \alpha_i \theta_i - \alpha_i z_i^\top x \right) + \frac{\lambda n}{2}\|x\|_2^2 \\
&\iff \min_x \sum_{i=1}^{n} \left(\min_{\theta_i} f_i(\theta_i) + \alpha_i \theta_i \right) + \frac{\lambda n}{2}\|x\|_2^2 - \sum_{i=1}^{n} \alpha_i z_i^\top x \\
&\iff -\sum_{i=1}^{n} f_i^*(-\alpha_i) + \min_x \frac{\lambda n}{2}\|x\|_2^2 - \sum_{i=1}^{n} \alpha_i z_i^\top x \\
&\iff \underbrace{-\sum_{i=1}^{n} f_i^*(-\alpha_i) - \frac{\lambda n}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i z_i \right\|_2^2}_{D(\alpha)} .
\end{aligned}$$

Step 3. Solve $\max_{\alpha \in \mathbb{R}^n} D(\alpha)$

3 Duality Gap

Recall from the previous section that the relation between primal variables and dual variables is

$$x = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i z_i.$$

The duality gap is defined by

$$\text{Duality gap} := F\big(x(\alpha)\big) - D(\alpha)$$

Then, the primal optimality gap $F(x(\alpha)) - F_*$ is bounded by the duality gap $:= F(x(\alpha)) - D(\alpha)$.

Remark: This reveals the benefit of considering developing algorithms in the dual space. Since we can obtain an upper-bound of the optimality gap on the fly during the execution of the underlying dual algorithm. We demonstrate one of the classical algorithms in the next section.

4 Stochastic Dual Coordinate Ascent (SDCA)

4.1 Main Idea

Consider the unconstrained optimization problem we introduced

$$\max_{\alpha \in \mathbb{R}^n} D(\alpha), \qquad \text{where } D(\alpha) := \frac{1}{n} \sum_{i=1}^{n} -f_i^*(-\alpha_i) - \frac{\lambda}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i z_i \right\|_2^2.$$

Consider updating a dual variable $\alpha_i \in \mathbb{R}^n$ at a time. That is, at the k-th iteration, we pick $i_k \in [n]$. Then, we have

$$\begin{aligned}
&\max_{\alpha_{i_k}} -\frac{1}{n} f_{i_k}^*\left(-\alpha_{i_k}\right) - \frac{\lambda}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i z_i \right\|_2^2 \\
\iff &\max_{\alpha_{i_k}} -\frac{1}{n} f_{i_k}^*\left(-\alpha_{i_k}\right) - \frac{\lambda}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i^{(k-1)} z_i + \frac{1}{\lambda n} \Delta\alpha_{i_k} z_{i_k} \right\|_2^2 \\
\iff &\max_{\Delta\alpha_{i_k}} -\frac{1}{n} f_{i_k}^*\left(-\left(\alpha_{i_k}^{(k-1)} + \Delta\alpha_{i_k}\right)\right) - \frac{\lambda}{2} \left\| x^{(k-1)} + \frac{1}{\lambda n} \Delta\alpha_{i_k} z_{i_k} \right\|_2^2,
\end{aligned}$$

where

$$\alpha_{i_k} = \underbrace{\alpha_{i_k}^{(k-1)}}_{\text{fixed}} + \underbrace{\Delta\alpha_{i_k}}_{\text{variable}}$$

and

$$x^{(k-1)} = \frac{1}{\lambda n}\sum_{i=1}^{n} \alpha_i^{(k-1)} z_i.$$

4.2 Algorithm

Below is a formal statement of the SDCA algorithm [3].

Algorithm 1 Stochastic Dual Coordinate Ascent (SDCA)

Init dual variables $\alpha^{(1)} \in \mathbb{R}^n$.

for $k = 1, 2, \ldots, K$ **do**

  Randomly pick a dual coordinate $i_k \in [n]$.

  Maximizes the dual problem by updating the dual variable i_k while fixing the others

$$\max_{\Delta\alpha_{i_k}} -\frac{1}{n} f_{i_k}^*\left(-\left(\alpha_{i_k}^{(k-1)} + \Delta\alpha_{i_k}\right)\right) - \frac{\lambda}{2}\left\| x^{(k-1)} + \frac{1}{\lambda n}\Delta\alpha_{i_k} z_{i_k}\right\|_2^2.$$

  $\alpha^{(k)} = \alpha^{(k-1)} + \Delta\alpha_{i_k} e_{i_k} \in \mathbb{R}^n$.

  $x^{(k)} = x^{(k-1)} + \frac{1}{\lambda n}\Delta\alpha_{i_k} z_{i_k} \in \mathbb{R}^d$.

end for

Output: $x(\alpha^{(K)}) := \frac{1}{\lambda n}\sum_{i=1}^{n} \alpha_i^{(K)} z_i$.

Remark: Note that in the primal space, each primal coordinate corresponds to a dimension of the "feature" vector; on the other hand, in the dual space, a dual coordinate corresponds to a data point. Randomly picking up a dual coordinate to update is about randomly choosing a sample to use for the update.

4.3 Example

Example Let us consider $f_i(\theta) := \max\{0, 1 - y_i\theta\}$ being the hinge loss, where $y_i \in \{-1, +1\}$. Its conjugate function is

$$f_i^*(a) = \begin{cases} ay_i & \text{, if } ay_i \in [-1, 0], \\ \infty & \text{, otherwise} \end{cases}.$$

The update of the SDCA for the hinge loss is

$$\Delta\alpha_{i_k} = y_{i_k} \max\left(0, \min\left(1, \frac{1 - z_{i_k}^\top x^{(k-1)} y_{i_k}}{\|z_{i_k}\|_2^2/\lambda n} + \alpha_{i_k}^{(k-1)} y_{i_k}\right)\right) - \alpha_{i_k}^{(k-1)}.$$

Bibliographic notes

For references on conjugate functions, please refer to Chapter 5 of Algorithms for Convex Optimization by Nisheeth K. Vishnoi [1] and Chapter 5 of Convex Optimization by Stephen Boyd and Lieven Vandenberghe.

References

[1] Nisheeth K. Vishnoi. Algorithms for Convex Optimization. Cambridge University Press, 2021.

[2] Stephen Boyd and Lieven Vandenberghe, Convex Optimization Cambridge University Press, 2004.

[3] Rie Johnson and Tong Zhang Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. NeurIPS 2013.