A Fenchel duality

Consider the optimization problem

$$\inf_{\mathbf{w}\in S} \left(f(\mathbf{w}) + \sum_{t=1}^{T} g_t(\mathbf{w}) \right) \ .$$

An equivalent problem is

$$\inf_{\mathbf{w}_0,\mathbf{w}_1,\ldots,\mathbf{w}_T} \left(f(\mathbf{w}_0) + \sum_{t=1}^{T} g_t(\mathbf{w}_t) \right) \ \text{s.t.}\ \ \mathbf{w}_0 \in S \ \text{ and } \ \forall t \in [T],\ \mathbf{w}_t = \mathbf{w}_0 \ . \tag{18}$$

Introducing T vectors $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_T$, each $\boldsymbol{\lambda}_t \in \mathbb{R}^n$ is a vector of Lagrange multipliers for the equality constraint $\mathbf{w}_t = \mathbf{w}_0$, we obtain the following Lagrangian

$$\mathcal{L}(\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_T, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_T) \ = \ f(\mathbf{w}_0) + \sum_{t=1}^{T} g_t(\mathbf{w}_t) + \sum_{t=1}^{T} \langle \boldsymbol{\lambda}_t, \mathbf{w}_0 - \mathbf{w}_t \rangle \ .$$

The dual problem is the task of maximizing the following dual objective value,

$$\begin{aligned} \mathcal{D}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_T) \ &= \ \inf_{\mathbf{w}_0 \in S, \mathbf{w}_1, \ldots, \mathbf{w}_T} \mathcal{L}(\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_T, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_T) \\ &= \ -\sup_{\mathbf{w}_0 \in S} \left(\left\langle \mathbf{w}_0, -\sum_{t=1}^{T} \boldsymbol{\lambda}_t \right\rangle - f(\mathbf{w}_0) \right) - \sum_{t=1}^{T} \sup_{\mathbf{w}_t} \left(\langle \mathbf{w}_t, \boldsymbol{\lambda}_t \rangle - g_t(\mathbf{w}_t) \right) \\ &= \ -f^\star \left(-\sum_{t=1}^{T} \boldsymbol{\lambda}_t \right) - \sum_{t=1}^{T} g_t^\star \left(\boldsymbol{\lambda}_t \right) \ , \end{aligned}$$

where $f^\star, g_1^\star, \ldots, g_T^\star$ are the Fenchel conjugate functions of $f, g_1, \ldots, g_T$. Therefore, the generalized Fenchel dual problem is

$$\sup_{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_T} \ - f^\star \left(-\sum_{t=1}^{T} \boldsymbol{\lambda}_t \right) - \sum_{t=1}^{T} g_t^\star(\boldsymbol{\lambda}_t) \ . \tag{19}$$

Note that when $T = 1$ the above duality is the so-called Fenchel duality [Borwein and Lewis, 2006].

The weak duality theorem tells us that the primal objective upper bounds the dual objective:

$$\sup_{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_T} -f^\star(-\sum_{t=1}^{T} \boldsymbol{\lambda}_t) - \sum_{t=1}^{T} g^\star(\boldsymbol{\lambda}_t) \ \leq \ \inf_{\mathbf{w} \in S} f(\mathbf{w}) + \sum_{t=1}^{T} g_t(\mathbf{w}) \, .$$

A sufficient condition for equality to hold (i.e. strong duality) is that f is a strongly convex function, $g_1, \ldots, g_T$ are convex functions, and the intersection of the domains of $g_1, \ldots, g_T$ is polyhedral.

Assume that strong duality holds, let $(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_T)$ be a maximizer of the dual objective function, and let $(\mathbf{w}_0^\star, \ldots, \mathbf{w}_T^\star)$ be a maximizer of the problem given in Eq. (18). Then, optimality conditions imply that

$$(\mathbf{w}_0^\star, \ldots, \mathbf{w}_T^\star) \ = \ \operatorname*{argmin}_{\mathbf{w}_0, \ldots, \mathbf{w}_T} \mathcal{L}(\mathbf{w}_0, \ldots, \mathbf{w}_T, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_T) \, .$$

See for example Boyd and Vandenberghe [2004] Section 5.5.2. In particular, the above implies that

$$\mathbf{w}_0^\star \ = \ \operatorname*{argmax}_{\mathbf{w}_0 \in S} \ \left\langle \mathbf{w}_0, -\sum_{t=1}^{T} \boldsymbol{\lambda}_t \right\rangle - f(\mathbf{w}_0) \ = \ \nabla f^\star(-\lambda_{1:T}) \, ,$$

where in the last equality we used Lemma 1. Since $\mathbf{w}_0^\star$ is also a minimizer of our original problem, we obtain the primal-dual link function

$$\mathbf{w} \ = \ \nabla f^\star(-\lambda_{1:T}) \, .$$

B Proof of Thm. 2

We first use the following properties of the Fenchel conjugate of strongly convex functions. The proof of this lemma follows from Lemma 18 in Shalev-Shwartz [2007].

Lemma 3 *Let f be a σ-strongly convex function over S with respect to a norm $\|\cdot\|$. Let $f^\star$ be the Fenchel conjugate of f. Then, $f^\star$ is differentiable and for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^n$, we have*

$$f^\star(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) - f^\star(\boldsymbol{\theta}_1) \ \leq\ \langle \nabla f^\star(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 \rangle + \frac{1}{2\sigma}\|\boldsymbol{\theta}_2\|_\star^2$$

We also need the following technical lemma.

Lemma 4 *Assume f strongly convex, let $a, b \geq 0$, and let $\mathbf{w}_b = \nabla f^\star(\boldsymbol{\theta}/b)$. Then,*

$$a f^\star(\boldsymbol{\theta}/a) - b f^\star(\boldsymbol{\theta}/b) \ \geq\ (b-a) f(\mathbf{w})$$

Proof Since f is strongly convex we know that $f^\star$ is differentiable. Using Lemma 1 we have

$$f^\star(\boldsymbol{\theta}/b) = \langle \mathbf{w}_b, \boldsymbol{\theta}/b \rangle - f(\mathbf{w}_b)$$

The definition of $f^\star$ now implies that

$$f^\star(\boldsymbol{\theta}/a) = \max_{\mathbf{w}} \langle \mathbf{w}, \boldsymbol{\theta}/a \rangle - f(\mathbf{w}) \geq \langle \mathbf{w}_b, \boldsymbol{\theta}/a \rangle - f(\mathbf{w}_b)$$

Therefore,

$$a f^\star(\boldsymbol{\theta}/a) - b f^\star(\boldsymbol{\theta}/b) \ \geq\ -(a-b) f(\mathbf{w}_b)$$

which concludes our proof. ■

Next, we show that the gradient descend update rule yields a sufficient increase of the dual objective.

Lemma 5 *Let $(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{t-1})$ be an arbitrary sequence of vectors. Denote $\mathbf{w} = \nabla f^\star\left(-\frac{1}{\sigma_{1:t}} \boldsymbol{\lambda}_{1:(t-1)}\right)$ and let $\boldsymbol{\lambda} \in \partial g_t(\mathbf{w})$. Then,*

$$\mathcal{D}_{t+1}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{t-1}, \boldsymbol{\lambda}) - \mathcal{D}_t(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{t-1}) \ \geq\ \ell_t(\mathbf{w}) - \frac{\|\boldsymbol{\lambda}\|_\star^2}{2\,\sigma_{1:t}}\ .$$

Proof Denote $\bar{\Delta}_t = \mathcal{D}_{t+1}(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{t-1}, \boldsymbol{\lambda}) - \mathcal{D}_t(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{t-1})$. Since f is strongly convex we can apply Lemma 3 to get that

$$\begin{aligned}
\bar{\Delta}_t &= -\sigma_{1:t} f^\star\left(-\tfrac{\boldsymbol{\lambda}_{1:(t-1)} + \boldsymbol{\lambda}}{\sigma_{1:t}}\right) + \sigma_{1:(t-1)} f^\star\left(-\tfrac{\boldsymbol{\lambda}_{1:(t-1)}}{\sigma_{1:(t-1)}}\right) - g_t^\star(\boldsymbol{\lambda}) \\
&\geq -\sigma_{1:t}\left(f^\star\left(-\tfrac{\boldsymbol{\lambda}_{1:(t-1)}}{\sigma_{1:t}}\right) - \tfrac{\langle \mathbf{w}, \boldsymbol{\lambda} \rangle}{\sigma_{1:t}} + \tfrac{\|\boldsymbol{\lambda}\|_\star^2}{2(\sigma_{1:t})^2}\right) + \sigma_{1:(t-1)} f^\star\left(-\tfrac{\boldsymbol{\lambda}_{1:(t-1)}}{\sigma_{1:(t-1)}}\right) - g_t^\star(\boldsymbol{\lambda}) \\
&= \underbrace{\sigma_{1:(t-1)} f^\star\left(-\tfrac{\boldsymbol{\lambda}_{1:(t-1)}}{\sigma_{1:(t-1)}}\right) - \sigma_{1:t} f^\star\left(-\tfrac{\boldsymbol{\lambda}_{1:(t-1)}}{\sigma_{1:t}}\right)}_{A} + \underbrace{\langle \mathbf{w}, \boldsymbol{\lambda} \rangle - g_t^\star(\boldsymbol{\lambda})}_{B} - \frac{\|\boldsymbol{\lambda}\|_\star^2}{2\sigma_{1:t}}\ .
\end{aligned}$$

Since $\boldsymbol{\lambda} \in \partial g_t(\mathbf{w})$ we get from Lemma 1 that $B = g_t(\mathbf{w})$. Next, we use Lemma 4 and the definition of $\mathbf{w}$ to get that $A \geq \left(\sigma_{1:t} - \sigma_{1:(t-1)}\right) f(\mathbf{w}) = \sigma_t f(\mathbf{w})$. Thus, $A + B \geq \sigma_t f(\mathbf{w}) + g_t(\mathbf{w}) = \ell_t(\mathbf{w})$ and this concludes our proof. ■

The proof of Thm. 2 now easily follows.

Proof [of Thm. 2] Denote $\Delta_t = \mathcal{D}_{t+1}(\boldsymbol{\lambda}_1^{t+1}, \ldots, \boldsymbol{\lambda}_t^{t+1}) - \mathcal{D}_t(\boldsymbol{\lambda}_1^t, \ldots, \boldsymbol{\lambda}_{t-1}^t)$ and note that Eq. (10) still holds. The definition of the update in Fig. 3 and Lemma 5 implies that there exists $\mathbf{v}_t \in \partial g_t(\mathbf{w}_t)$ such that $\Delta_t \geq \ell_t(\mathbf{w}_t) - \frac{\|\mathbf{v}_t\|_\star^2}{2\sigma_{1:t}}$. Summing over t and combining with Eq. (10) we conclude our proof. ■