1 Convex duality

We define the dual (conjugate) of $f$ as

$$f^*(v) = \sup_u [u^T v - f(u)].$$

Note that $f^*(v)$ is convex (even if $f(u)$ isn’t because it is the sup of convex functions).

By definition, we have the following inequality:

$$u^T v \leq f(u) + f^*(v),$$

which decouples $u$ and $v$.

We also have $f(u) = (f^*)^*(u)$. To see this, we use the following (not exactly rigorous derivation): given any $u$, let $v_0 = \nabla f(u)$, then

$$u^T v_0 = f(u) + f^*(v_0)$$

because $u^T v_0 - f(u)$ (which is concave) has subgradient zero at $u$, and thus achieves maximum. Now, we know that

$$(f^*)^*(u) \geq u^T v_0 - f^*(v_0)$$

and thus $(f^*)^*(u) \geq f(u)$.

In addition, we know that there exists $v_0'$ such that

$$(f^*)^*(u) + f^*(v_0') - u^T v_0' = 0.$$  

This means that $f(u) \geq u^T v_0' - f^*(v_0') = (f^*)^*(u)$. Therefore we have $f(u) = (f^*)^*(u)$. Note that if $(u, v)$ is a pair such that the equality holds $u^T v = f(u) + f^*(v)$, then we have the relationship $u = \nabla f^*(v)$ and $v = \nabla f(u)$.

Some examples of convex duality (verification leaves as exercise):

$$f(u) = p^{-1} \|u\|_p^p; \quad f^*(v) = q^{-1} \|v\|_q^q \quad (p^{-1} + q^{-1} = 1).$$

$$f(u) = 0.5 \|u\|_p^2; \quad f^*(v) = 0.5 \|v\|_q^2 \quad (p^{-1} + q^{-1} = 1).$$

If $\sum_j \mu_j = 1$ and $\mu_j \geq 0$, then

$$f(u) = \ln \sum_j \mu_j e^{u_j}; \quad f^*(v) = \sum_j v_j \ln(v_j/\mu_j) \text{ subject to } \sum_j v_j = 1, v_j \geq 0.$$  

For any norm $\|u\|_p$, one can also define its dual norm $\|v\|_D$ as

$$\|v\|_D = \sup_{\|u\| \leq 1} u^T v.$$
This means that we have the decoupling inequality:

\[ u^\top v \leq \|u\|_p \|v\|_D. \]

Examples: for vectors, \(\|u\|_p\) and \(\|v\|_q\) are dual norms when \(1/p + 1/q = 1\). The same holds for matrix Schatten norms.

## 2 Rademacher Complexity of Regularized Linear Function Class

Consider linear functions of the form:

\[ F = \{ w^\top x : g(w) \leq A \}, \]

and we are interested in its Rademacher Complexity:

\[ R_n(F, X_1^n) = E_\sigma \sup_{\|w\|_p \leq A} n^{-1} \sum_{i=1}^n \sigma_i w^\top X_i. \]

Then using duality, we have

\[ n^{-1} \sum_{i=1}^n \sigma_i w^\top X_i \leq \inf_{\lambda} [\lambda^{-1} g(w) + \lambda^{-1} g^*(\lambda n^{-1} \sum_{i=1}^n \sigma_i X_i)]. \]

If \(g^*(0) = 0\) and is smooth with respect to a norm \(\|\cdot\|\):

\[ g^*(u) \leq g^*(v) + \nabla g^*(v)^\top (u - v) + L\|u - v\|^2, \]

for some \(L > 0\), then one can show using induction that

\[ R_n(F, X_1^n) \leq \inf_{\lambda} [\lambda^{-1} g(w) + \lambda^{-1} E_\sigma g^*(\lambda n^{-1} \sum_{i=1}^n \sigma_i X_i)] \]

\[ \leq \inf_{\lambda} \left[ \lambda^{-1} g(w) + \lambda^{-1} 0.5 E_{\sigma_i^{-1}} [g^*(\lambda n^{-1} (-X_n + \sum_{i=1}^{n-1} \sigma_i X_i)) + g^*(\lambda n^{-1} (X_n + \sum_{i=1}^{n-1} \sigma_i X_i))] \right] \]

\[ \leq \inf_{\lambda} \left[ \lambda^{-1} g(w) + \lambda n^{-2} \|X_n\|^2 + \lambda^{-1} E_{\sigma_i^{-1}} g^*(\lambda n^{-1} (\sum_{i=1}^{n-1} \sigma_i X_i)) \right] \]

\[ \cdots \]

\[ \leq 2 \sqrt{ALn^{-2} \sum_{i=1}^n \|X_i\|^2}. \]

Then

\[ R_n(F, X_1^n) \leq 2 \sqrt{ABL/n}, \quad B = \frac{1}{n} \sum_{i=1}^n \|X_i\|^2. \]

## 3 Some Examples

### 3.1 Vector \(L_2\) regularization

We have \(g(w) = 0.5\|w\|_2^2\), then \(g^*(u) = 0.5\|u\|_2^2\), and is smooth with respect to \(\|\cdot\|_2\) with \(L = 0.5\). It follows that

\[ R_n(F, X_1^n) \leq ab/\sqrt{n}; \quad F = \{ w : \|w\|_2 \leq a \}; \quad b = \sup_i \|X_i\|_2. \]
3.2 Vector $L_p$ regularization

We have $g(w) = 0.5\|w\|^2_p$ with $p \in (1, 2]$, then $g^*(w) = 0.5\|u\|^2_p$, where $1/p + 1/q = 1$. It can be shown with Taylor expansion that $g^*(\cdot)$ is smooth with respect to $\|\cdot\|_q$ with $L = 0.5(q - 1)$. It follows that

$$R_n(F, X^n_\mu) \leq ab\sqrt{(q - 1)/n}; \quad F = \{w : \|w\|_p \leq a\}; \quad b = \sup_i \|X_i\|_q.$$  

Note that this formula diverges when $p = 1$ (corresponding to $q = \infty$). We need another formulation to deal with the case $p = 1$ (or $p$ is close to 1).

Note that $w$ can be infinite dimensional.

3.3 Vector entropy regularization

Here we assume that constraint that $\sum_j w_j = A_1$ and $w_j \geq 0$ (note that we can transform $x \to [x, -x]$ to simulate the effect of $w_j \leq 0$). In this case, we consider regularization $g(w) = \sum_j w_j \ln(\mu_j/w_j)$, where $\{\mu_j > 0\}$ is a set of positive prior such that $\sum_j \mu_j = A_1$. In this case, we know that $g^*(u) = A_1 \ln(\sum_j (\mu_j/A_1) \exp(u_j))$, and $g^*(u)$ is smooth with respect to $\|\cdot\|_\infty$ with $L = 0.5 A_1$. It follows that

$$R_n(F, X^n_\mu) \leq B\sqrt{2A_1 A_2}/n; \quad F = \{w : \sum_j w_j \ln(\mu_j/w_j) \leq A_2; w_j \geq 0; \sum_j w_j = A_1\}; \quad B = \sup_i \|X_i\|_\infty.$$  

Here $w$ can be infinite dimensional. In finite dimension, where $w, x \in \mathbb{R}^p$, we may take $\mu_j = A_1/p$, and the maximum value $\sum_j w_j \ln(\mu_j/w_j) \leq A_1 \ln(p)$. Therefore we may take $A_2 = A_1 \ln(p)$ and obtain the following bound for $L_1$ regularization (in finite dimension):

$$R_n(F, X^n_\mu) \leq A_1 b\sqrt{2\ln(p)}/n; \quad F = \{w : w_j \geq 0; \sum_j w_j = A_1\}; \quad b = \sup_i \|X_i\|_\infty.$$  

3.4 Matrix $L_p$ Schatten norm regularization

Let $w$ be a matrix, and $g(w) = 0.5\|w\|^2_p$ with $p \in (1, 2]$, where $\|\cdot\|_p$ denotes the matrix Schatten norm here. Then the results essentially follow that of the vector norm, with $g^*(u) = 0.5\|u\|^2_p$, where $1/p + 1/q = 1$. It can be shown with Taylor expansion that $g^*(\cdot)$ is smooth with respect to $\|\cdot\|_q$ with $L = 0.5(q - 1)$. It follows that

$$R_n(F, X^n_\mu) \leq ab\sqrt{(q - 1)/n}; \quad F = \{w : \|w\|_p \leq a\}; \quad b = \sup_i \|X_i\|_q.$$  

Similar results parallel to vector entropy regularization can be obtained for matrix regularization.