Appendix

Appendix (References)

[CN07] Castro & Nowak, (2007) Minimax Bounds for Active Learning. COLT 2007

[HK11] Hazan & Kale (2011) Beyond The Regret Minimization Barrier: An Optimal Algorithm for Stochastic Strongly-Convex Optimization. *COLT 2011*

Section 2

Lemma 1. No function can satisfy Uniform Convexity for $\kappa < 2$, but they can be in \mathcal{F}^{κ} for $\kappa < 2$.

Proof. If uniform convexity could be satisfied for (say) $\kappa = 1.5$, then we have for all $x, y \in S$

$$f(y) - f(x) - g_x^{\top}(y - x) \ge \frac{\lambda}{2} ||x - y||_2^{1.5}$$

Take x, y both on the positive **x**-axis. The Taylor expansion would require, for some $c \in [x, y]$,

$$f(y) - f(x) - g_x^{\top}(y - x) = \frac{1}{2}(x - y)^{\top} H(c)(x - y)$$
$$\leq \frac{\|H(c)\|_F}{2} \|x - y\|_2^2$$

Now, taking $||x-y||_2 = \epsilon \to 0$ by choosing x closer to y, the Taylor condition requires the residual to grow like ϵ^2 (going to zero fast), but the UC condition requires the residual to grow at least as fast as $\epsilon^{1.5}$ (going to zero slow). At some small enough value of ϵ , this would not be possible. Since the definition of UC needs to hold for all $x, y \in S$, this gives us a contradiction. So, no f can be uniformly convex for any $\kappa < 2$

However, one can note that for $f(x) = ||x||_{1.5}^{1.5} = \sum_{i} |x_{i}|^{1.5}$, we have $x_{f}^{*} = 0$, and $f(x) - f(x_{f}^{*}) = ||x||_{1.5}^{1.5} \ge ||x - x_{f}^{*}||_{2}^{1.5}$, hence $f \in \mathcal{F}^{1.5}$.

Lemma 2. If $f \in \mathcal{F}^{\kappa}$, then for any subgradient $g_x \in \partial f(x)$, we have $||g_x||_2 \ge \lambda ||x - x^*||_2^{\kappa-1}$.

Proof. By convexity, we have

$$f(x^*) \ge f(x) + g_x^{\top}(x^* - x)$$

Rearranging terms and since $f \in \mathcal{F}^{\kappa}$, we get

$$g_x^{\top}(x - x^*) \ge f(x) - f(x^*) \ge \lambda ||x - x^*||_2^{\kappa}$$

By Holder's inequality,

$$||g_x||_2 ||x - x^*||_2 \ge g_x^+ (x - x^*)$$

Putting them together, we have

$$||g_x||_2 ||x - x^*||_2 \ge \lambda ||x - x^*||_2^{\kappa}$$

giving us our result.

Lemma 3. For a gaussian random variable z, $\forall t < \sigma$, $\exists a_1, a_2, a_1t \leq P(0 \leq z \leq t) \leq a_2t$

Proof. We wish to characterize how the probability mass of a gaussian random variable grows just around its mean. Our claim is that it grows linearly with the distance from the mean, and the following simple argument argues this neatly.

Consider a $X \sim N(0, \sigma^2)$ random variable at a distance t from the mean 0. We want to bound $\int_{-t}^{t} d\mu(X)$ for very small t. The key idea in bounding this integral is to approximate it by a smaller and larger rectangle, each of the rectangles having a width 2t (from -t to t).

The first one has a height equal to $\frac{e^{-t^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$, the smallest value taken by the gaussian in [-t, t] achieved at t, and the other with a height equal to the $\frac{1}{\sigma\sqrt{2\pi}}$, the largest value of the gaussian in [-t, t] achieved at 1.

The smaller rectangle has area $2t \frac{e^{-t^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \geq 2t \frac{e^{-1/2}}{\sigma\sqrt{2\pi}}$ when $t < \sigma$. The larger rectangle clearly has an area of $2t \frac{1}{\sigma\sqrt{2\pi}}$.

Hence we have $A_1 t = 2t \frac{1}{\sigma\sqrt{2\pi e}} \leq P(|X| < t) \leq 2t \frac{1}{\sigma\sqrt{2\pi}} = A_2 t$ for $t < \sigma$. Similarly, for a one-sided inequality, we have $a_1 t = t \frac{1}{\sigma\sqrt{2\pi e}} \leq P(0 < X < t) \leq t \frac{1}{\sigma\sqrt{2\pi}} = a_2 t$ for $t < \sigma$.

We note that the gaussian tail inequality $P(X > t) \leq \frac{1}{t}e^{-t^2/2\sigma^2}$ really makes sense for large $t > \sigma$ and we are interested in $t < \sigma$. There are tighter inequalities,

but for our purpose, this will suffice.

Lemma 4. If $|\eta(x) - 1/2| \ge \lambda$, the midpoint \hat{x}_T of the high-probability interval returned by BZ satisfies $\mathbb{E}|\hat{x}_T - x^*| = O(e^{-T\lambda^2/2})$. [CN07]

Proof. The BZ algorithm works by dividing [0, 1] into a grid of m points (interval size 1/m) and makes Tqueries (only at gridpoints) to return an interval \hat{I}_T such that $\Pr(x^* \notin \hat{I}_T) \leq m e^{-T\lambda^2}$ [CN07]. We choose \hat{x}_T to be the midpoint of this interval, and hence get

$$\mathbb{E}|\hat{x}_{T} - x^{*}| = \int_{0}^{1} \Pr(|\hat{x}_{T} - x^{*}| > u) du$$

=
$$\int_{0}^{1/2m} \Pr(|\hat{x}_{T} - x^{*}| > u) du$$

+
$$\int_{1/2m}^{1} \Pr(|\hat{x}_{T} - x^{*}| > u) du$$

$$\leq \frac{1}{2m} + \left(1 - \frac{1}{2m}\right) \Pr\left(|\hat{x}_{T} - x^{*}| > \frac{1}{2m}\right)$$

$$\leq \frac{1}{2m} + me^{-T\lambda^{2}} = O\left(e^{-T\lambda^{2}/2}\right)$$

for the choice of the number of gridpoints as $m = e^{T\lambda^2/2}$.

Lemma 5. If $|\eta(x) - 1/2| \ge \lambda |x - x^*|^{\kappa}$, the point \hat{x}_T obtained from a modified version of BZ satisfies $\mathbb{E}|\hat{x}_T - x^*| = O\left(\left(\frac{\log T}{T}\right)^{\frac{1}{2\kappa-2}}\right)$ and $\mathbb{E}[|\hat{x}_T - x^*|^{\kappa}] = O\left(\left(\frac{\log T}{T}\right)^{\frac{\kappa}{2\kappa-2}}\right)$.

Proof. We again follow the same proof as in [CN07]. Initially, they assume that the grid points are not aligned with x^* , ie $\forall k \in \{0, ..., m\}, \quad |x^* - k/m| \geq 1/3m$. This implies that for all gridpoints $x, |\eta(x) - 1/2| \geq \lambda(1/3m)^{\kappa-1}$. Following the exact same proof above,

$$\mathbb{E}[|\hat{x}_{T} - x^{*}|^{\kappa}] = \int_{0}^{1} \Pr(|\hat{x}_{T} - x^{*}|^{\kappa} > u) du$$

$$= \int_{0}^{(1/2m)^{\kappa}} \Pr(|\hat{x}_{T} - x^{*}| > u^{1/\kappa}) du$$

$$+ \int_{(1/2m)^{\kappa}}^{1} \Pr(|\hat{x}_{T} - x^{*}| > u^{1/\kappa}) du$$

$$\leq \left(\frac{1}{2m}\right)^{\kappa} + \left(1 - \left(\frac{1}{2m}\right)^{\kappa}\right) \Pr\left(|\hat{x}_{T} - x^{*}| > \frac{1}{2m}\right)^{\kappa}$$

$$\leq \left(\frac{1}{2m}\right)^{\kappa} + m \exp(-T\lambda^{2}(1/3m)^{2\kappa-2})$$

$$= O\left(\left(\frac{T}{\log T}\right)^{\frac{1}{2\kappa-2}}\right)$$

on choosing *m* proportional to $\left(\frac{T}{\log T}\right)^{\frac{1}{2\kappa-2}}$.

[CN07] elaborate in detail how to avoid the assumption that the grid points don't align with x^* . They use a more complicated variant of BZ with three interlocked grids, and gets the same rate as above without that assumption. The reader is directed to their exposition for clarification.

Section 3

Lemma 6. $c_{\kappa} \|x\|_{\kappa}^{\kappa} = c_{\kappa} \sum_{i=1}^{d} |x_i|^{\kappa} =: f_0(x) \in \mathcal{F}^{\kappa}$, for all $\kappa > 1$. Also, $f_1(x)$ as defined in Section 3 is also in \mathcal{F}^{κ} .

Proof. Firstly, this is clearly convex for $\kappa > 1$. Also, $f_0(x_{f_0}^*) = 0$ at $x_{f_0}^* = 0$. So, all we need to show is that for appropriate choice of c_{κ} , f is indeed 1-Lipschitz and that $f_0(x) - f_0(x_{f_0}^*) \ge \lambda ||x - x_{f_0}^*||_2^{\kappa}$ for some $\lambda > 0$, ie

$$c_{\kappa} \|x\|_{\kappa}^{\kappa} \ge \lambda \|x\|_{2}^{\kappa} \quad , \quad c_{\kappa}(\|x\|_{\kappa}^{\kappa} - \|y\|_{\kappa}^{\kappa}) \le \|x - y\|_{2}$$

Let us consider two cases, $\kappa \geq 2$ and $\kappa < 2$. Note that all norms are uniformly bounded with respect to each other, upto constants depending on d. Precisely, if $\kappa < 2$, then $||x||_{\kappa} > ||x||_2$ and if $\kappa \geq 2$, then $||x||_{\kappa} \geq d^{1/\kappa - 1/2} ||x||_2$.

When $\kappa \geq 2$, consider $c_{\kappa} = 1$. Then

$$(\|x\|_{\kappa}^{\kappa} - \|y\|_{\kappa}^{\kappa}) \le \|x - y\|_{\kappa}^{\kappa} \le \|x - y\|_{2}^{\kappa} \le \|x - y\|_{2}$$

because $||z||_{\kappa} \leq ||z||_2$ and $||x - y|| \leq 1$. Also, $||x||_{\kappa}^{\kappa} \geq d^{1-\frac{\kappa}{2}} ||x||_2^{\kappa}$, so $\lambda = d^{1-\frac{\kappa}{2}}$ works.

When $\kappa < 2$, consider $c_{\kappa} = \frac{1}{\sqrt{d^{\kappa}}}$. Similarly

$$c_{\kappa}(\|x\|_{\kappa}^{\kappa} - \|y\|_{\kappa}^{\kappa}) \le \left(\frac{\|x-y\|_{\kappa}}{\sqrt{d}}\right)^{\kappa} \le \|x-y\|_{2}^{\kappa} \le \|x-y\|_{2}$$

Also $c_{\kappa} \|x\|_{\kappa}^{\kappa} \ge c_{\kappa} \|x\|_{2}^{\kappa}$, so $\lambda = c_{\kappa}$ works.

Hence $f_0(x)$ is 1-Lipschitz and in \mathcal{F}^{κ} for appropriate c_{κ} .

Now, look at $f_1(x)$ for $x_1 \leq 4a$. It is actually just $f_0(x)$, but translated by 2a in direction x_1 , with a constant added, and hence has the same growth around its minimum. Now, the part with $x_1 > 4a$ is just $f_0(x)$ itself, which have the same growth parameters as the part with $x_1 \leq 4a$. So $f_1(x) \in \mathcal{F}^{\kappa}$ also.

Lemma 7. For all i = 1...d, let $f_i(x)$ be any onedimensional κ -uniformly convex function ($\kappa \ge 2$) with constant λ_i . For a d-dimensional function $f(x) = \sum_{i=1}^d f_i(x_i)$ that decomposes over dimensions, f(x) is also κ -uniformly convex with constant $\lambda = \frac{\min_i \lambda_i}{d^{1/2-1/\kappa}}$.

Proof.

$$f(x+h) = \sum_{i} f_{i}(x_{i}+h_{i})$$

$$\geq \sum_{i} (f_{i}(x_{i}) + g_{x_{i}}h_{i} + \lambda_{i}|h_{i}|^{\kappa})$$

$$\geq f(x) + g_{x}^{\top}h + (\min_{i}\lambda_{i})||h||_{\kappa}^{\kappa}$$

$$\geq f(x) + g_{x}^{\top}h + \frac{(\min_{i}\lambda_{i})}{d^{1/2-1/\kappa}}||h||_{2}^{\kappa}$$

(one can use h = y - x for the usual first-order definition)

Lemma 8. $f(x) = |x|^k$ is κ -uniformly convex i.e.

$$tf(x) + (1-t)f(y) \ge f(tx + (1-t)y) + \frac{\lambda}{2}t(1-t)|x-y|^k$$

for $\lambda = 4/2^k$. Lemma 7 implies $||x||_{\kappa}^{\kappa}$ is also κ -uniformly convex with $\lambda = \frac{4/2^k}{d^{1/2-1/\kappa}}$.

Proof. First we will show this for the special case of t = 1/2. We need to argue that:

$$\frac{1}{2}|x|^k + \frac{1}{2}|y|^k \ge |\frac{x+y}{2}|^k + \lambda \frac{1}{8}|x-y|^k$$

Let $\lambda = 4/2^k$. We will prove a stronger claim -

$$\frac{1}{2}|x|^k + \frac{1}{2}|y|^k \ge |\frac{x+y}{2}|^k + 2\lambda \frac{1}{8}|x-y|^k$$

Since $k \geq 2$

$$\begin{aligned} RHS^{1/k} &= (|\frac{x+y}{2}|^k + |\frac{x-y}{2}|^k)^{1/k} \\ &\leq (|\frac{x+y}{2}|^2 + |\frac{x-y}{2}|^2)^{1/2} \\ &\leq (|x|^2/2 + |y|^2/2)^{1/2} \\ &\leq \frac{1}{\sqrt{2}} 2^{1/2 - 1/k} (|x|^k + |y|^k)^{1/k} \\ &\leq (\frac{1}{2}|x|^k + \frac{1}{2}|y|^k)^{1/k} = LHS^{1/k} \end{aligned}$$

Now, for the general case. We will argue that just proving the above for t = 1/2 is actually sufficient.

$$\begin{aligned} f(tx + (1-t)y) &= f\left(2t\left(\frac{x+y}{2}\right) + (1-2t)y\right) \\ &\leq & 2tf\left(\frac{x+y}{2}\right) + (1-2t)f(y) \\ &\leq & tf(x) + tf(y) - 2t\frac{2\lambda}{8}|x-y|^k + (1-2t)f(y) \\ &\leq & tf(x) + (1-t)f(y) - t(1-t)\frac{\lambda}{2}|x-y|^k \end{aligned}$$