## 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 $\mathcal{F}^{\kappa}$ for $\kappa<2$.

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

$$
f(y)-f(x)-g_{x}^{\top}(y-x) \geq \frac{\lambda}{2}\|x-y\|_{2}^{1.5}
$$

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

$$
\begin{gathered}
f(y)-f(x)-g_{x}^{\top}(y-x)=\frac{1}{2}(x-y)^{\top} H(c)(x-y) \\
\leq \frac{\|H(c)\|_{F}}{2}\|x-y\|_{2}^{2}
\end{gathered}
$$

Now, taking $\|x-y\|_{2}=\epsilon \rightarrow 0$ by choosing $x$ closer to $y$, the Taylor condition requires the residual to grow like $\epsilon^{2}$ (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 $\epsilon$, 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 $f(x)=\|x\|_{1.5}^{1.5}=$ $\sum_{i}\left|x_{i}\right|^{1.5}$, we have $x_{f}^{*}=0$, and $f(x)-f\left(x_{f}^{*}\right)=$ $\|x\|_{1.5}^{1.5} \geq\left\|x-x_{f}^{*}\right\|_{2}^{1.5}$, 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 $\left\|g_{x}\right\|_{2} \geq \lambda\left\|x-x^{*}\right\|_{2}^{\kappa-1}$.

Proof. By convexity, we have

$$
f\left(x^{*}\right) \geq f(x)+g_{x}^{\top}\left(x^{*}-x\right)
$$

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

$$
g_{x}^{\top}\left(x-x^{*}\right) \geq f(x)-f\left(x^{*}\right) \geq \lambda\left\|x-x^{*}\right\|_{2}^{\kappa}
$$

By Holder's inequality,

$$
\left\|g_{x}\right\|_{2}\left\|x-x^{*}\right\|_{2} \geq g_{x}^{\top}\left(x-x^{*}\right)
$$

Putting them together, we have

$$
\left\|g_{x}\right\|_{2}\left\|x-x^{*}\right\|_{2} \geq \lambda\left\|x-x^{*}\right\|_{2}^{\kappa}
$$

giving us our result.

Lemma 3. For a gaussian random variable $z, \forall t<$ $\sigma, \quad \exists a_{1}, a_{2}, \quad a_{1} t \leq P(0 \leq z \leq t) \leq a_{2} t$

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 $X \sim N\left(0, \sigma^{2}\right)$ 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 $2 t$ (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 $2 t \frac{e^{-t^{2} / 2 \sigma^{2}}}{\sigma \sqrt{2 \pi}} \geq 2 t \frac{e^{-1 / 2}}{\sigma \sqrt{2 \pi}}$ when $t<\sigma$. The larger rectangle clearly has an area of $2 t \frac{1}{\sigma \sqrt{2 \pi}}$.
Hence we have $A_{1} t=2 t \frac{1}{\sigma \sqrt{2 \pi e}} \leq P(|X|<t) \leq$ $2 t \frac{1}{\sigma \sqrt{2 \pi}}=A_{2} t$ 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 1
but for our purpose, this will suffice.

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

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

$$
\begin{aligned}
& \mathbb{E}\left|\hat{x}_{T}-x^{*}\right|=\int_{0}^{1} \operatorname{Pr}\left(\left|\hat{x}_{T}-x^{*}\right|>u\right) d u \\
= & \int_{0}^{1 / 2 m} \operatorname{Pr}\left(\left|\hat{x}_{T}-x^{*}\right|>u\right) d u \\
& +\int_{1 / 2 m}^{1} \operatorname{Pr}\left(\left|\hat{x}_{T}-x^{*}\right|>u\right) d u \\
\leq & \frac{1}{2 m}+\left(1-\frac{1}{2 m}\right) \operatorname{Pr}\left(\left|\hat{x}_{T}-x^{*}\right|>\frac{1}{2 m}\right) \\
\leq & \frac{1}{2 m}+m e^{-T \lambda^{2}}=O\left(e^{-T \lambda^{2} / 2}\right)
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
& \mathbb{E}\left[\left|\hat{x}_{T}-x^{*}\right|^{\kappa}\right]=\int_{0}^{1} \operatorname{Pr}\left(\left|\hat{x}_{T}-x^{*}\right|^{\kappa}>u\right) d u \\
= & \int_{0}^{(1 / 2 m)^{\kappa}} \operatorname{Pr}\left(\left|\hat{x}_{T}-x^{*}\right|>u^{1 / \kappa}\right) d u \\
& +\int_{(1 / 2 m)^{\kappa}}^{1} \operatorname{Pr}\left(\left|\hat{x}_{T}-x^{*}\right|>u^{1 / \kappa}\right) d u \\
\leq & \left(\frac{1}{2 m}\right)^{\kappa}+\left(1-\left(\frac{1}{2 m}\right)^{\kappa}\right) \operatorname{Pr}\left(\left|\hat{x}_{T}-x^{*}\right|>\frac{1}{2 m}\right) \\
\leq & \left(\frac{1}{2 m}\right)^{\kappa}+m \exp \left(-T \lambda^{2}(1 / 3 m)^{2 \kappa-2}\right)
\end{aligned}
$$

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

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. $c_{\kappa}\|x\|_{\kappa}^{\kappa}=c_{\kappa} \sum_{i=1}^{d}\left|x_{i}\right|^{\kappa}=: f_{0}(x) \in \mathcal{F}^{\kappa}$, for all $\kappa>1$. Also, $f_{1}(x)$ as defined in Section 3 is also in $\mathcal{F}^{\kappa}$.

Proof. Firstly, this is clearly convex for $\kappa>1$. Also, $f_{0}\left(x_{f_{0}}^{*}\right)=0$ at $x_{f_{0}}^{*}=0$. So, all we need to show is that for appropriate choice of $c_{\kappa}, f$ is indeed 1-Lipschitz and that $f_{0}(x)-f_{0}\left(x_{f_{0}}^{*}\right) \geq \lambda\left\|x-x_{f_{0}}^{*}\right\|_{2}^{\kappa}$ for some $\lambda>0$, ie

$$
c_{\kappa}\|x\|_{\kappa}^{\kappa} \geq \lambda\|x\|_{2}^{\kappa} \quad, \quad c_{\kappa}\left(\|x\|_{\kappa}^{\kappa}-\|y\|_{\kappa}^{\kappa}\right) \leq\|x-y\|_{2}
$$

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

$$
\left(\|x\|_{\kappa}^{\kappa}-\|y\|_{\kappa}^{\kappa}\right) \leq\|x-y\|_{\kappa}^{\kappa} \leq\|x-y\|_{2}^{\kappa} \leq\|x-y\|_{2}
$$

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
$c_{\kappa}\left(\|x\|_{\kappa}^{\kappa}-\|y\|_{\kappa}^{\kappa}\right) \leq\left(\frac{\|x-y\|_{\kappa}}{\sqrt{d}}\right)^{\kappa} \leq\|x-y\|_{2}^{\kappa} \leq\|x-y\|_{2}$
Also $c_{\kappa}\|x\|_{\kappa}^{\kappa} \geq c_{\kappa}\|x\|_{2}^{\kappa}$, so $\lambda=c_{\kappa}$ works.
Hence $f_{0}(x)$ is 1-Lipschitz and in $\mathcal{F}^{\kappa}$ for appropriate $c_{\kappa}$.

Now, look at $f_{1}(x)$ for $x_{1} \leq 4 a$. It is actually just $f_{0}(x)$, but translated by $2 a$ in direction $x_{1}$, with a constant added, and hence has the same growth around its minimum. Now, the part with $x_{1}>4 a$ is just $f_{0}(x)$ itself, which have the same growth parameters as the part with $x_{1} \leq 4 a$. So $f_{1}(x) \in \mathcal{F}^{\kappa}$ also.

Lemma 7. For all $i=1 \ldots d$, let $f_{i}(x)$ be any onedimensional $\kappa$-uniformly convex function ( $\kappa \geq 2$ ) with constant $\lambda_{i}$. For a d-dimensional function $f(x)=$ $\sum_{i=1}^{d} f_{i}\left(x_{i}\right)$ that decomposes over dimensions, $f(x)$ is also $\kappa$-uniformly convex with constant $\lambda=\frac{\min _{i} \lambda_{i}}{d^{1 / 2-1 / \kappa}}$.

Proof.

$$
\begin{aligned}
& f(x+h)=\sum_{i} f_{i}\left(x_{i}+h_{i}\right) \\
\geq & \sum_{i}\left(f_{i}\left(x_{i}\right)+g_{x_{i}} h_{i}+\lambda_{i}\left|h_{i}\right|^{\kappa}\right) \\
\geq & f(x)+g_{x}^{\top} h+\left(\min _{i} \lambda_{i}\right)\|h\|_{\kappa}^{\kappa} \\
\geq & f(x)+g_{x}^{\top} h+\frac{\left(\min _{i} \lambda_{i}\right)}{d^{1 / 2-1 / \kappa}}\|h\|_{2}^{\kappa}
\end{aligned}
$$

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

Lemma 8. $f(x)=|x|^{k}$ is $\kappa$-uniformly convex i.e.
$t f(x)+(1-t) f(y) \geq f(t x+(1-t) y)+\frac{\lambda}{2} t(1-t)|x-y|^{k}$
for $\lambda=4 / 2^{k}$. Lemma 7 implies $\|x\|_{\kappa}^{\kappa}$ is also $\kappa$ 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:

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

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

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

Since $k \geq 2$

$$
\begin{aligned}
R H S^{1 / k} & =\left(\left|\frac{x+y}{2}\right|^{k}+\left|\frac{x-y}{2}\right|^{k}\right)^{1 / k} \\
& \leq\left(\left|\frac{x+y}{2}\right|^{2}+\left|\frac{x-y}{2}\right|^{2}\right)^{1 / 2} \\
& \leq\left(|x|^{2} / 2+|y|^{2} / 2\right)^{1 / 2} \\
& \leq \frac{1}{\sqrt{2}} 2^{1 / 2-1 / k}\left(|x|^{k}+|y|^{k}\right)^{1 / k} \\
& \leq\left(\frac{1}{2}|x|^{k}+\frac{1}{2}|y|^{k}\right)^{1 / k}=L H S^{1 / k}
\end{aligned}
$$

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

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

