Supplementary material to "Online A-Optimal Design and Active Linear Regression"

Xavier Fontaine [1] **Pierre Perrault** [2] **Michal Valko** [3] **Vianney Perchet** [4 5]

A. Concentration arguments

In this section we present results on the concentration of the variance for subgaussian random variables. Traditional results on the concentration of the variances (Maurer & Pontil, 2009; Carpentier et al., 2011) are obtained in the bounded setting. We propose results in a more general framework. Let us begin with some definitions.

Definition S1 (Sub-gaussian random variable)**.** *A random variable X is said to be κ^2-sub-gaussian if*

$$\forall \lambda \geq 0,\ \exp(\lambda(X - \mathbb{E}X)) \leq \exp(\lambda^2\kappa^2/2)\ .$$

And we define its ψ_2-norm as

$$\|X\|_{\psi_2} = \inf\left\{t > 0 \,|\, \mathbb{E}[\exp(X^2/t^2)] \leq 2\right\}\ .$$

We can bound the ψ_2-norm of a subgaussian random variable as stated in the following lemma.

Lemma S1 (ψ_2-norm)**.** *If X is a centered κ^2-sub-gaussian random variable then*

$$\|X\|_{\psi_2} \leq \frac{2\sqrt{2}}{\sqrt{3}}\kappa\ .$$

Proof. A proposition from (Wainwright, 2019) shows that for all $\lambda \in [0, 1)$, a sub-gaussian variable X verifies

$$\mathbb{E}\left(\frac{\lambda X^2}{2\kappa^2}\right) \leq \frac{1}{\sqrt{1-\lambda}}\ .$$

Taking $\lambda = 3/4$ and defining $u = \frac{2\sqrt{2}}{\sqrt{3}}\kappa$ gives

$$\mathbb{E}(X^2/u^2) \leq 2\ .$$

Consequently $\|X\|_{\psi_2} \leq u$. □

A wider class of random variables is the class of sub-exponential random variables that are defined as follows.

Definition S2 (Sub-exponential random variable)**.** *A random variable X is said to be sub-exponential if there exists $K > 0$ such that*

$$\forall\, 0 \leq \lambda \leq 1/K,\ \mathbb{E}[\exp(\lambda|X|)] \leq \exp(K\lambda)\ .$$

And we define its ψ_1-norm as

$$\|X\|_{\psi_1} = \inf\left\{t > 0 \,|\, \mathbb{E}[\exp(|X|/t)] \leq 2\right\}\ .$$

A result from (Vershynin, 2018) gives the following lemma, that makes a connection between subgaussian and subexponential random variables.

[1]Centre Borelli, ENS Paris-Saclay, Palaiseau, France [2]Idemia, Courbevoie, France [3]Google DeepMind, Paris, France [4]CREST, ENSAE, Palaiseau, France [5]Criteo AI Lab, Paris, France. Correspondence to: Xavier Fontaine <xavier.fontaine@polytechnique.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Lemma S2. *A random variable X is sub-gaussian if and only if X^2 is sub-exponential, and we have*

$$\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2 \ .$$

We now want to obtain a concentration inequality on the empirical variance of a sub-gaussian random variable. We give use the following notations to define the empirical variance.

Definition S3. *We define the following quantities for n i.i.d repetitions of the random variable X.*

$$\mu = \mathbb{E}[X] \qquad \textit{and} \qquad \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i \ ,$$
$$\mu^{(2)} = \mathbb{E}[X^2] \qquad \textit{and} \qquad \hat{\mu}_n^{(2)} = \frac{1}{n}\sum_{i=1}^n X_i^2 \ .$$

The variance and empirical variance are defined as follows

$$\sigma^2 = \mu^{(2)} - \mu^2 \qquad \textit{and} \qquad \hat{\sigma}_n^2 = \hat{\mu}_n^{(2)} - \hat{\mu}_n^2 \ .$$

We are now able to prove Theorem 1 that we restate below for clarity.

Theorem S1. *Let X be a centered and κ^2-sub-gaussian random variable sampled $n \geq 2$ times. Let $\delta \in (0,1)$. Let $c = (e-1)(2e(2e-1))^{-1} \approx 0.07$. With probability at least $1-\delta$, the following concentration bound on its empirical variance hold*

$$\left|\hat{\sigma}_n^2 - \sigma^2\right| \leq \frac{8}{3}\kappa^2 \cdot \max\left(\frac{\log(4/\delta)}{cn}, \sqrt{\frac{\log(4/\delta)}{cn}}\right) + 2\kappa^2\frac{\log(4/\delta)}{n} \ .$$

Proof. We have

$$\begin{aligned}
\left|\hat{\sigma}_n^2 - \sigma^2\right| &= \left|\hat{\mu}_n^{(2)} - \hat{\mu}_n^2 - (\mu^{(2)} - \mu^2)\right| \\
&\leq \left|\hat{\mu}_n^{(2)} - \mu^{(2)}\right| + \left|\hat{\mu}_n^2 - \mu^2\right| \\
&\leq \left|\hat{\mu}_n^{(2)} - \mu^{(2)}\right| + |\hat{\mu}_n - \mu||\hat{\mu}_n + \mu| \\
&\leq \left|\hat{\mu}_n^{(2)} - \mu^{(2)}\right| + |\hat{\mu}_n|^2
\end{aligned}$$

since $\mu = 0$.

We now apply Hoeffding's inequality to the X_t variables that are κ^2-subgaussian, to get

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu > t\right) \leq \exp\left(-\frac{n^2t^2}{2n\kappa^2}\right) = \exp\left(-\frac{nt^2}{2\kappa^2}\right).$$

And finally

$$\mathbb{P}\left(|\hat{\mu}_n - \mu| > \kappa\sqrt{\frac{2\log(2/\delta)}{n}}\right) \leq \delta.$$

Consequently with probability at least $1-\delta$, $|\hat{\mu}_n|^2 \leq 2\kappa^2\frac{\log(2/\delta)}{n}$.

The variables X_t^2 are sub-exponential random variables. We can apply Bernstein's inequality as stated in (Chafaï et al., 2012) to get for all $t > 0$:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i^2 - \mu^{(2)}\right| > t\right) \leq 2\exp\left(-cn\min\left(\frac{t^2}{s^2}, \frac{t}{m}\right)\right)$$

$$\leq 2\exp\left(-cn\min\left(\frac{t^2}{m^2}, \frac{t}{m}\right)\right) .$$

with $c = \frac{e-1}{2e(2e-1)}$, $s^2 = \frac{1}{n}\sum_{i=1}^n \left\|X_i^2\right\|_{\psi_1} \leq m^2$ and $m = \max_{1\leq i\leq n}\left\|X_i^2\right\|_{\psi_1}$.

Inverting the inequality we obtain

$$\mathbb{P}\left(\left|\hat{\mu}_n^{(2)} - \mu^{(2)}\right| > m\cdot\max\left(\frac{\log(2/\delta)}{cn}, \sqrt{\frac{\log(2/\delta)}{cn}}\right)\right) \leq \delta .$$

And finally, with probability at least $1-\delta$,

$$\left|\hat{\sigma}_n^2 - \sigma^2\right| \leq m\cdot\max\left(\frac{\log(4/\delta)}{cn}, \sqrt{\frac{\log(4/\delta)}{cn}}\right) + 2\kappa^2\frac{\log(4/\delta)}{n} .$$

Using Lemmas S2 and S1 we obtain that $m \leq 8\kappa^2/3$. Finally,

$$\begin{aligned}\left|\hat{\sigma}_n^2 - \sigma^2\right| &\leq \frac{8}{3}\kappa^2\cdot\max\left(\frac{\log(4/\delta)}{cn}, \sqrt{\frac{\log(4/\delta)}{cn}}\right) + 2c\kappa^2\frac{\log(4/\delta)}{cn} \\ &\leq 3\kappa^2\cdot\max\left(\frac{\log(4/\delta)}{cn}, \sqrt{\frac{\log(4/\delta)}{cn}}\right) ,\end{aligned}$$

since $2c \leq 1/3$. This gives the expected result. □

We now state a corollary of this result.

Corollary 1. *Let $T \geq 2$. Let X be a centered and κ^2-sub-gaussian random variable. Let $c = (e-1)(2e(2e-1))^{-1} \approx 0.07$. For $n = \left\lceil \frac{72\kappa^4}{c\sigma^4}\log(2T)\right\rceil$, we have with probability at least $1-1/T^2$,*

$$\left|\hat{\sigma}_n^2 - \sigma^2\right| \leq \frac{1}{2}\sigma^2.$$

Proof. Let $\delta \in (0,1)$. Let $n = \left\lceil \frac{\log(4/\delta)}{c}\left(\frac{6\kappa^2}{\sigma^2}\right)^2\right\rceil$.

Then $\frac{\log(4/\delta)}{cn} \leq \left(\frac{\sigma^2}{6\kappa^2}\right)^2 < 1$, since $\sigma^2 \leq \kappa^2$, by property of subgaussian random variables.

With probability $1-\delta$, Theorem 1 gives

$$|\hat{\sigma}_n^2 - \sigma^2| \leq 3\kappa^2\frac{\sigma^2}{6\kappa^2} \leq \frac{1}{2}\sigma^2 .$$

Now, suppose that $\delta = 1/T^2$. Then, with probability $1-1/T^2$, for $n = \left\lceil \frac{72\kappa^4}{c\sigma^4}\log(2T)\right\rceil$ samples,

$$|\hat{\sigma}_n^2 - \sigma^2| \leq \frac{1}{2}\sigma^2 .$$

□

B. Proof of gradient concentration

In this section we prove Proposition 3.

Proof. Let $p \in \Delta^K$ and let $i \in [K]$. We compute

$$\begin{aligned} G_i - \hat{G}_i &= \left\| \hat{\Omega}(p)^{-1} \frac{X_i}{\hat{\sigma}_i} \right\|_2^2 - \left\| \Omega(p)^{-1} \frac{X_i}{\sigma_i} \right\|_2^2 \\ &\leq \left\| \hat{\Omega}(p)^{-1} \frac{X_i}{\hat{\sigma}_i} - \Omega(p)^{-1} \frac{X_i}{\sigma_i} \right\|_2 \left\| \hat{\Omega}(p)^{-1} \frac{X_i}{\hat{\sigma}_i} + \Omega(p)^{-1} \frac{X_i}{\sigma_i} \right\|_2 . \end{aligned}$$

Let us now note $A \doteq \hat{\Omega}(p)\hat{\sigma}_i$ and $B \doteq \Omega(p)\sigma_i$. We have, using that $\|X_k\|_2 = 1$,

$$\begin{aligned} \left\| \hat{\Omega}(p)^{-1} \frac{X_k}{\hat{\sigma}_k} - \Omega(p)^{-1} \frac{X_k}{\sigma_k} \right\|_2 &= \left\| (A^{-1} - B^{-1}) X_k \right\|_2 \\ &\leq \left\| A^{-1} - B^{-1} \right\|_2 \|X_k\|_2 \\ &\leq \left\| A^{-1}(B - A)B^{-1} \right\|_2 \\ &\leq \left\| A^{-1} \right\|_2 \left\| B^{-1} \right\|_2 \|B - A\|_2 . \end{aligned}$$

One of the quantity to bound is $\left\| B^{-1} \right\|_2$. We have

$$\left\| B^{-1} \right\|_2 = \rho(B^{-1}) = \frac{1}{\min(\mathrm{Sp}(B))} ,$$

where $\mathrm{Sp}(B)$ is the spectrum (set of eigenvalues) of B. We know that $\mathrm{Sp}(B) = \sigma_i \mathrm{Sp}(\Omega(p))$. Therefore we need to find the smallest eigenvalue λ of $\Omega(p)$. Since the matrix is invertible we know $\lambda > 0$.

We will need the following lemma.

Lemma S3. *Let $\mathbb{X}_0 = \left(X_1^\top, \cdots, X_k^\top \right)^\top$. We have*

$$\lambda_{\min}(\Omega(p)) \geq \min_{k \in [K]} \frac{p_k}{\sigma_k^2} \lambda_{\min}(\mathbb{X}_0^\top \mathbb{X}_0).$$

Proof. We have for all $p \in \Delta^K$,

$$\min_{i \in [K]} \frac{p_i}{\sigma_i^2} \sum_{k=1}^K X_k X_k^\top \preccurlyeq \sum_{k=1}^K \frac{p_k}{\sigma_k^2} X_k X_k^\top .$$

Therefore

$$\min_{k \in [K]} \frac{p_k}{\sigma_k^2} \mathbb{X}_0^\top \mathbb{X}_0 \preccurlyeq \Omega(p) .$$

And finally

$$\min_{k \in [K]} \frac{p_k}{\sigma_k^2} \lambda_{\min}(\mathbb{X}_0^\top \mathbb{X}_0) \leq \lambda_{\min}(\Omega(p)) .$$

□

Note now that the smallest eigenvalue of $\mathbb{X}_0^\top \mathbb{X}_0$ is actually the smallest non-zero eigenvalue of $\mathbb{X}_0 \mathbb{X}_0^\top$, which is the Gram matrix of $(X_1, \ldots, X_d)$, that we note now Γ. This directly gives the following

Proposition S1.

$$\left\| B^{-1} \right\|_2 \leq \frac{1}{\sigma_i \lambda_{\min}(\Gamma)} \max_{k \in [K]} \frac{\sigma_k^2}{p_k}.$$

We jump now to the bound of $\left\| A^{-1} \right\|_2$. We could obtain a similar bound to the one of $\left\| B^{-1} \right\|_2$ but it would contain $\hat{\sigma}_k$ values. Since we do not want a bound containing estimates of the variances, we prove the

Proposition S2.

$$\left\| A^{-1} \right\|_2 \leq 2 \left\| B^{-1} \right\|_2 .$$

Proof. We have, if we note $H = A - B$,

$$\left\|A^{-1}\right\|_2 = \left\|(B + A - B)^{-1}\right\|_2 \leq \left\|B^{-1}\right\|_2 \left\|(I_n + B^{-1}H)^{-1}\right\|_2 \leq 2\left\|B^{-1}\right\|_2 \ ,$$

from a certain rank. □

Let us now bound $\|B - A\|_2$. We have

$$\begin{aligned}
\|B - A\|_2 &= \left\| \sigma_i \sum_{k=1}^K p_k \frac{X_k X_k^\top}{\sigma_k^2} - \hat{\sigma}_i \sum_{k=1}^K p_k \frac{X_k X_k^\top}{\hat{\sigma}_k^2} \right\|_2 \\
&= \left\| \sum_{k=1}^K p_k X_k X_k^\top \left(\frac{\sigma_i}{\sigma_k^2} - \frac{\hat{\sigma}_i}{\hat{\sigma}_k^2} \right) \right\|_2 \\
&\leq \sum_{k=1}^K p_k \left| \frac{\sigma_i}{\sigma_k^2} - \frac{\hat{\sigma}_i}{\hat{\sigma}_k^2} \right| \|X_k\|_2^2 \\
&\leq \sum_{k=1}^K p_k \left| \frac{\sigma_i}{\sigma_k^2} - \frac{\hat{\sigma}_i}{\hat{\sigma}_k^2} \right|.
\end{aligned}$$

The next step is now to use Theorem 1 in order to bound the difference $\left| \frac{\sigma_i}{\sigma_k^2} - \frac{\hat{\sigma}_i}{\hat{\sigma}_k^2} \right|$.

Proposition S3. *With the notations introduced above, we have*

$$\|B - A\|_2 \leq \frac{113 K \sigma_{\max}}{\sigma_{\min}^4} \kappa_{\max}^2 \cdot \max\left(\frac{\log(4TK/\delta)}{T_i}, \sqrt{\frac{\log(4TK/\delta)}{T_i}} \right) \ .$$

Proof. Corollary 1 gives that for all $k \in [K]$, $\frac{1}{2}\sigma_k^2 \leq \hat{\sigma}_k^2 \leq \frac{3}{2}\sigma_k^2$.

A consequence of Theorem 1 is that for all $k \in [K]$, if we note T_k the (random) number of samples of covariate k, we have, with probability at least $1 - \delta$,

$$\forall k \in [K], \left|\sigma_k^2 - \hat{\sigma}_k^2\right| \leq \frac{8}{3}\kappa_k^2 \cdot \max\left(\frac{\log(4TK/\delta)}{cT_k}, \sqrt{\frac{\log(4TK/\delta)}{cT_k}} \right) + 2\kappa_k^2 \frac{\log(4TK/\delta)}{T_k}.$$

We note Δ_k the r.h.s of the last equation. We begin by establishing a simple upper bound of Δ_k. Using the fact that $\sqrt{1/c} \leq 1/c$ and that $8/(3c) \leq 38$, we have

$$\begin{aligned}
\Delta_k &\leq \frac{8}{3c}\kappa_k^2 \cdot \max\left(\frac{\log(4TK/\delta)}{T_k}, \sqrt{\frac{\log(4TK/\delta)}{T_k}} \right) + 2\kappa_k^2 \frac{\log(4TK/\delta)}{T_k} \\
&\leq 38\kappa_k^2 \cdot \max\left(\frac{\log(4TK/\delta)}{T_k}, \sqrt{\frac{\log(4TK/\delta)}{T_k}} \right) + 2\kappa_k^2 \frac{\log(4TK/\delta)}{T_k} \\
&\leq 40\kappa_k^2 \cdot \max\left(\frac{\log(4TK/\delta)}{T_k}, \sqrt{\frac{\log(4TK/\delta)}{T_k}} \right).
\end{aligned}$$

Let $k \in [K]$. We have

$$\left| \frac{\sigma_i}{\sigma_k^2} - \frac{\hat{\sigma}_i}{\hat{\sigma}_k^2} \right| = \left| \frac{\sigma_i \hat{\sigma}_k^2 - \hat{\sigma}_i \sigma_k^2}{\sigma_k^2 \hat{\sigma}_k^2} \right| = \left| \frac{\sigma_i \hat{\sigma}_k^2 - \sigma_i \sigma_k^2 + \sigma_i \sigma_k^2 - \hat{\sigma}_i \sigma_k^2}{\sigma_k^2 \hat{\sigma}_k^2} \right|$$

$$
\begin{aligned}
&\leq \left|\frac{\sigma_i(\hat{\sigma}_k^2-\sigma_k^2)}{\sigma_k^2\hat{\sigma}_k^2}\right| + \left|\frac{\sigma_i-\hat{\sigma}_i}{\hat{\sigma}_k^2}\right| \\
&\leq \left|\frac{\sigma_i(\hat{\sigma}_k^2-\sigma_k^2)}{\sigma_k^2\hat{\sigma}_k^2}\right| + \left|\frac{\sigma_i^2-\hat{\sigma}_i^2}{\hat{\sigma}_k^2(\sigma_i+\hat{\sigma}_i)}\right| \\
&\leq \left|\frac{\sigma_i(\hat{\sigma}_k^2-\sigma_k^2)}{\sigma_k^2\hat{\sigma}_k^2}\right| + \left|\frac{\sigma_i^2-\hat{\sigma}_i^2}{\hat{\sigma}_k^2\sigma_i}\right| \\
&\leq \left|\hat{\sigma}_k^2-\sigma_k^2\right|\left|\frac{\sigma_i}{\sigma_k^2\hat{\sigma}_k^2}\right| + \left|\sigma_i^2-\hat{\sigma}_i^2\right|\left|\frac{1}{\hat{\sigma}_k^2\sigma_i}\right| \\
&\leq \Delta_k\frac{2\sigma_{\max}}{\sigma_{\min}^4} + \Delta_i\frac{2\sqrt{2}}{\sigma_{\min}^3}.
\end{aligned}
$$

Finally we have, using the fact that $T \geq T_k$ for all $k \in [K]$

$$
\begin{aligned}
\|B-A\|_2 &\leq \sum_{k=1}^K p_k\left|\frac{\sigma_i}{\sigma_k^2}-\frac{\hat{\sigma}_i}{\hat{\sigma}_k^2}\right| \\
&\leq \frac{2\sigma_{\max}}{\sigma_{\min}^4}\left(\sum_{k=1}^K p_k\Delta_k + \sqrt{2}\sum_{k=1}^K p_k\Delta_i\right) \\
&\leq \frac{2\sigma_{\max}}{\sigma_{\min}^4}\left(\sum_{k=1}^K \frac{T_k}{T}40\kappa_k^2\cdot\max\left(\frac{\log(4TK/\delta)}{T_k}, \sqrt{\frac{\log(4TK/\delta)}{T_k}}\right) + \sqrt{2}\Delta_i\right) \\
&\leq \frac{2\sigma_{\max}}{\sigma_{\min}^4}\left(\sum_{k=1}^K 40\kappa_k^2\cdot\max\left(\frac{\log(4TK/\delta)}{T}, \sqrt{\frac{T_k}{T}}\sqrt{\frac{\log(4TK/\delta)}{T}}\right) + \sqrt{2}\Delta_i\right) \\
&\leq \frac{2\sigma_{\max}}{\sigma_{\min}^4}\left(\sum_{k=1}^K 40\kappa_k^2\cdot\max\left(\frac{\log(4TK/\delta)}{T}, \sqrt{\frac{\log(4TK/\delta)}{T}}\right) + \sqrt{2}\Delta_i\right) \\
&\leq \frac{2\sigma_{\max}}{\sigma_{\min}^4}\left(K40\kappa_{\max}^2\cdot\max\left(\frac{\log(4TK/\delta)}{T_i}, \sqrt{\frac{\log(4TK/\delta)}{T_i}}\right) + \sqrt{2}\Delta_i\right) \\
&\leq (K+\sqrt{2})\frac{80\sigma_{\max}}{\sigma_{\min}^4}\kappa_{\max}^2\cdot\max\left(\frac{\log(4TK/\delta)}{T_i}, \sqrt{\frac{\log(4TK/\delta)}{T_i}}\right).
\end{aligned}
$$

□

The last quantity to bound to end the proof is $\left\|\hat{\Omega}(p)^{-1}\frac{X_k}{\hat{\sigma}_k} + \Omega(p)^{-1}\frac{X_k}{\sigma_k}\right\|_2$.

Proposition S4. *We have*

$$
\left\|\hat{\Omega}(p)^{-1}\frac{X_k}{\hat{\sigma}_k} + \Omega(p)^{-1}\frac{X_k}{\sigma_k}\right\|_2 \leq 3\left\|B^{-1}\right\|_2.
$$

Proof. We have

$$
\begin{aligned}
\left\|\hat{\Omega}(p)^{-1}\frac{X_k}{\hat{\sigma}_k} + \Omega(p)^{-1}\frac{X_k}{\sigma_k}\right\|_2 &= \left\|(A^{-1}+B^{-1})X_k\right\|_2 \\
&\leq \left\|A^{-1}+B^{-1}\right\|_2\|X_k\|_2 \\
&\leq \left\|(A^{-1}-B^{-1})+2B^{-1}\right\|_2 \\
&\leq \left\|A^{-1}-B^{-1}\right\|_2 + 2\left\|B^{-1}\right\|_2.
\end{aligned}
$$

For T sufficiently large we have $\left\|\hat{\Omega}(p)^{-1}\frac{X_k}{\hat{\sigma}_k} + \Omega(p)^{-1}\frac{X_k}{\sigma_k}\right\|_2 \leq 3\left\|B^{-1}\right\|_2$. □

Combining Propositions S1, S2, S3 and S4 we obtain that $G_i - \hat{G}_i \leq 6 \left\|B^{-1}\right\|_2^3 \|B - A\|_2$ and

$$G_i - \hat{G}_i \leq 678K \frac{\sigma_{\max}}{\sigma_{\min}^4} \left(\frac{1}{\sigma_i \lambda_{\min}(\Gamma)} \max_{k \in [K]} \frac{\sigma_k^2}{p_k} \right)^3 \cdot \kappa_{\max}^2 \cdot \max\left(\frac{\log(4TK/\delta)}{T_i}, \sqrt{\frac{\log(4TK/\delta)}{T_i}} \right) ,$$

which proves Proposition 3.

□

C. Proofs of preliminary and easy results

In all the following we will denote by $\preccurlyeq$ the Loewner ordering: if A and B are two symmetric matrices, $A \preccurlyeq B$ iff $B - A$ is positive semi-definite.

C.1. Proof of Proposition 1

Proof. Let $p, q \in \mathring{\Delta}^d$, so that $\Omega(p)$ and $\Omega(q)$ are invertible, and $\lambda \in [0, 1]$. We have $L(p) = \mathrm{Tr}(\Omega(p)^{-1})$ and $L(\lambda p + (1 - \lambda)q) = \mathrm{Tr}(\Omega(\lambda p + (1 - \lambda)q)^{-1})$, where

$$\begin{aligned} \Omega(\lambda p + (1 - \lambda q)) &= \sum_{k=1}^{d} \frac{\lambda p_k + (1 - \lambda) q_k}{\sigma_k^2} X_k X_k^\top \\ &= \lambda \Omega(p) + (1 - \lambda)\Omega(q). \end{aligned}$$

It is well-known (Whittle, 1958) that the inversion is strictly convex on the set of positive definite matrices. Consequently,

$$\Omega(\lambda p + (1 - \lambda q))^{-1} = (\lambda \Omega(p) + (1 - \lambda)\Omega(q))^{-1} \prec \lambda \Omega(p)^{-1} + (1 - \lambda)\Omega(q)^{-1}.$$

Taking the trace this gives

$$L(\lambda p + (1 - \lambda)q) < \lambda L(p) + (1 - \lambda)L(q).$$

Hence L is convex. □

C.2. Proof of Lemma S4

Lemma S4. *Let S be a symmetric positive definite matrix and D a diagonal matrix with strictly positive entries $d_1, \ldots, d_n$. Then*

$$\lambda_{\min}(DSD) \geq \min_i (d_i)^2 \lambda_{\min}(S).$$

Proof. We have $\lambda_{\min}(S) Id \preccurlyeq S$ and consequently, multiplying by D (positive definite) to the right and left we obtain $\lambda_{\min}(S) D^2 \preccurlyeq DSD$, hence

$$\min_i (d_i)^2 \lambda_{\min}(S) \leq \lambda_{\min}(DSD).$$

□

D. Proofs of the slow rates

D.1. Proof of Proposition 2

Proof. We now conduct the analysis of Algorithm 1. Our strategy will be to convert the error $L(p_T) - L(p^\star)$ into a sum over $t \in [T]$ of small errors. Notice first that the quantity

$$\left\|\Omega(p)^{-1} X_k\right\|_2^2$$

can be upper bounded by $\frac{1}{\sigma_i \lambda_{\min}(G)} \max_{k \in [K]} \frac{\sigma_k^2}{0.5 p^o}$, for $p = p_T$. For $p = \hat{p}_t$, we can also bound this quantity by $\frac{4}{\sigma_i \lambda_{\min}(G)} \max_{k \in [K]} \frac{\sigma_k^2}{0.5 p^o}$, using Lemma 3 to express $\hat{p}_t$ with respect to lower estimates of the variances — and thus with respect to real variance thanks to Corollary 1. Then, from the convexity of L, we have

$$L(p_T) - L(p^\star) = L(p_T) - L\left(1/T\sum_{t=1}^T \hat{p}_t\right) + L\left(\frac{1}{T}\sum_{t=1}^T \hat{p}_t\right) - L(p^\star)$$
$$\leq \sum_k - \left\|\Omega(p_T)^{-1}\frac{X_k}{\sigma_k}\right\|_2^2 \left(p_{k,T} - \frac{1}{T}\sum_{t=1}^T \hat{p}_{k,t}\right) + \frac{1}{T}\sum_{t=1}^T \left(L(\hat{p}_t) - L(p^\star)\right)$$

Using Hoeffding inequality, $\left(p_{k,T} - \frac{1}{T}\sum_{t=1}^T \hat{p}_{k,t}\right) = \frac{1}{T}\sum_{t=1}^T \left(\mathbb{I}\{k \text{ is sampled at } t\} - \hat{p}_{k,t}\right)$ is bounded by $\sqrt{\frac{\log(2/\delta)}{T}}$ with probability $1-\delta$. It thus remains to bound the second term $\frac{1}{T}\sum_{t=1}^T \left(L(\hat{p}_t) - L(p^\star)\right)$. First, notice that $L(p)$ is an increasing function of σ_i for any i. If we define $\hat{L}$ be replacing each σ_i^2 by lower confidence estimates of the variances $\tilde{\sigma}_i^2$ (see Theorem 1), then

$$L(\hat{p}_t) - L(p^\star) \leq L(\hat{p}_t) - \hat{L}(p^\star) = L(\hat{p}_t) - \hat{L}(\hat{p}_t) + \hat{L}(\hat{p}_t) - \hat{L}(p^*) \leq L(\hat{p}_t) - \hat{L}(\hat{p}_t).$$

Since the gradient of L with respect to σ^2 is $\left(\frac{2p_i}{\sigma_i^3}\left\|\Omega(p)^{-1}X_i\right\|_2^2\right)_i$, we can bound $L(\hat{p}_t) - \hat{L}(\hat{p}_t)$ by

$$1/\sigma_{\min}^3 \sup_k \left\|\Omega(\hat{p}_t)^{-1}X_k\right\|_2^2 \sum_i 2\hat{p}_{i,t}|\sigma_i^2 - \tilde{\sigma}_i^2|.$$

Since $\hat{p}_{i,t}$ is the probability of having a feedback from covariate i, we can use the probabilistically triggered arm setting of Wang & Chen (2017) to prove that $\frac{1}{T}\sum_{t=1}^T \sum_i 2\hat{p}_i|\sigma_i^2 - \tilde{\sigma}_i^2| = \mathcal{O}\left(\sqrt{\frac{\log(T)}{T}}\right)$. Taking δ of order T^{-1} gives the desired result. □

E. Analysis of the bandit algorithm

E.1. Proof of Lemma 1

We begin by a lemma giving the coefficients of $\Omega(p)^{-1}$.

Lemma S5. *The diagonal coefficients of $\Omega(p)^{-1}$ can be computed as follows:*

$$\forall i \in [d],\ \Omega(p)_{ii}^{-1} = \sum_{j=1}^d \frac{\sigma_j^2 \operatorname{Cof}(\mathbb{X}_0^\top)_{ij}^2}{\det(\mathbb{X}_0^T\mathbb{X}_0)}\frac{1}{p_j}\ .$$

Proof. We suppose that $\forall i \in [d],\ p_i \neq 0$ so that $\Omega(p)$ is invertible.

We know that $\Omega(p)^{-1} = \frac{\operatorname{Com}(\Omega(p))^\top}{\det(\Omega(p))}$. We compute now $\det(\Omega(p))$.

$$\det(\Omega(p)) = \det\left(\sum_{k=1}^d \frac{p_k X_k X_k^\top}{\sigma_k^2}\right) = \det((\sqrt{T^{-1}}\mathbb{X})^\top\sqrt{T^{-1}}\mathbb{X}) = T^{-d}\det(\mathbb{X}^\top)^2$$
$$= T^{-d}\begin{vmatrix} & \vdots & \\ \tilde{X}_1 & \vdots & \tilde{X}_d \\ & \vdots & \end{vmatrix}^2 = \begin{vmatrix} & \vdots & \\ \frac{\sqrt{p_1}}{\sigma_1}X_1 & \vdots & \frac{\sqrt{p_d}}{\sigma_d}X_d \\ & \vdots & \end{vmatrix}^2$$
$$= \det(\mathbb{X}_0)^2\frac{p_1}{\sigma_1^2}\cdots\frac{p_d}{\sigma_d^2}\ .$$

We now compute $\mathrm{Com}(\Omega(p))_{ii}$.

$$\mathrm{Com}(\Omega(p)) = \mathrm{Com}(T^{-1/2}\mathbb{X}^\top T^{-1/2}\mathbb{X}) = \mathrm{Com}(T^{-1/2}\mathbb{X}^\top)\,\mathrm{Com}(T^{-1/2}\mathbb{X}^\top)^\top\ .$$

Let us note $M \doteq T^{-1/2}\mathbb{X} = \begin{pmatrix} \dots & \frac{\sqrt{p_1}}{\sigma_1}X_1^\top & \dots \\ & \vdots & \\ \dots & \frac{\sqrt{p_K}}{\sigma_K}X_K^\top & \dots \end{pmatrix}$. Therefore

$$\mathrm{Com}(\Omega(p))_{ii} = \sum_{j=1}^d \mathrm{Com}(M^\top)_{ij}^2 = \sum_{j=1}^d \prod_{k\neq j}\frac{p_k}{\sigma_k^2}\,\mathrm{Cof}(\mathbb{X}_0^\top)_{ij}^2\ .$$

Finally,

$$\Omega(p)_{ii}^{-1} = \sum_{j=1}^d \frac{\sigma_j^2\,\mathrm{Cof}(\mathbb{X}_0^\top)_{ij}^2}{\det(\mathbb{X}_0^\top\mathbb{X}_0)}\frac{1}{p_j}\ .$$

□

This allows us to derive the exact expression of the loss function L and we restate Lemma 1.

Lemma S6. *We have, for all $p \in \Delta^d$,*

$$L(p) = \frac{1}{\det(\mathbb{X}_0^\top\mathbb{X}_0)}\sum_{k=1}^d \frac{\sigma_k^2}{p_k}\,\mathrm{Cof}(\mathbb{X}_0\mathbb{X}_0^\top)_{kk}\ .$$

Proof. Using Lemma S5 we obtain

$$\begin{aligned} L(p) &= \mathrm{Tr}(\Omega(p)^{-1}) = \sum_{k=1}^d \Omega(p)_{kk}^{-1} \\ &= \frac{1}{\det(\mathbb{X}^\top\mathbb{X})}\sum_{k=1}^d\frac{\sigma_k^2}{p_k}\sum_{i=1}^d \mathrm{Cof}(\mathbb{X}_0^\top)_{ik}^2 = \frac{1}{\det(\mathbb{X}_0^\top\mathbb{X}_0)}\sum_{k=1}^d\frac{\sigma_k^2}{p_k}\,\mathrm{Com}(\mathbb{X}_0\mathbb{X}_0^\top)_{kk}\ . \end{aligned}$$

□

E.2. Proof of Lemma 4

Proof. We use the fact that for all $i \in [d]$, $p_i \geq p_i^o/2$. We have that for all $i \in [d]$,

$$\nabla_{ii}^2 L(p) = \frac{\mathrm{Cof}(\Gamma)_{ii}\sigma_i^2}{\det(\Gamma)}\frac{2}{p_i^3} \leq \frac{2\ \mathrm{Cof}(\Gamma)_{ii}\sigma_i^2}{\det(\Gamma)(p_i^o/2)^3}.$$

We have $p_k^o = \dfrac{\overline{\sigma}_k\sqrt{\mathrm{Cof}(\Gamma)_{kk}}}{\sum_{i=1}^d \overline{\sigma}_i\sqrt{\mathrm{Cof}(\Gamma)_{ii}}}$ which gives

$$\nabla_{ii}^2 L(p) \leq 16\frac{\sigma_{\max}^2\left(\sum_{k=1}^d \overline{\sigma}_k\sqrt{\mathrm{Cof}(\Gamma)_{kk}}\right)^3}{\det(\Gamma)\overline{\sigma}_{\min}^3\sqrt{\min_k \mathrm{Cof}(\Gamma)_{kk}}} \doteq C_S.$$

And consequently L is C_S-Lipschitz smooth.

We can obtain an upper bound on C_S using Corollary 1, which tells that $\sigma_k/2 \leq \overline{\sigma}_k \leq 3\sigma_k/2$:

$$C_S \leq 432\frac{\sigma_{\max}^2\left(\sum_{k=1}^d \sigma_k\sqrt{\mathrm{Cof}(\Gamma)_{kk}}\right)^3}{\det(\Gamma)\sigma_{\min}^3\sqrt{\min_k \mathrm{Cof}(\Gamma)_{kk}}}.$$

□

E.3. Proof of Theorem 2

Proof. Proposition 3 gives that

$$|G_i - \hat{G}_i| \leq 678K \frac{\sigma_{\max}}{\sigma_{\min}^4} \left(\frac{1}{\sigma_i \lambda_{\min}(\text{Gram})} \max_{k \in [K]} \frac{\sigma_k^2}{p_k} \right)^3 \cdot \kappa_{\max}^2 \cdot \max \left(\frac{\log(4TK/\delta)}{T_i}, \sqrt{\frac{\log(4TK/\delta)}{T_i}} \right).$$

Since each arm has been sampled at least a linear number of times we guarantee that $\log(4TK/\delta)/T_i \leq 1$ such that

$$|G_i - \hat{G}_i| \leq 678K \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^7 \frac{1}{\lambda_{\min}(\Gamma)^3} \frac{\kappa_{\max}^2}{p_{\min}^3} \sqrt{\frac{\log(4TK/\delta)}{T_i}}.$$

Thanks to the presampling phase of Lemma 3, we know that $p_{\min} \geq p^o/2$. For the sake of clarity we note $C \doteq 678K \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^7 \frac{8}{{p^o}^3 \lambda_{\min}(\Gamma)^3} \kappa_{\max}^2$ such that $|G_i - \hat{G}_i| \leq C \sqrt{\frac{\log(4TK/\delta)}{T_i}}$.

We have seen that L is μ-strongly convex, C_L-smooth and that $\text{dist}(p^\star, \partial \Delta^d) \geq \eta$. Consequently, since Lemma 3 shows that the pre-sampling stage does not affect the convergence result, we can apply (Berthet & Perchet, 2017, Theorem 7) (with the choice $\delta_T = 1/T^2$, which gives that

$$\mathbb{E}[L(p_T)] - L(p^\star) \leq c_1 \frac{\log^2(T)}{T} + c_2 \frac{\log(T)}{T} + c_3 \frac{1}{T},$$

with $c_1 = \frac{96C^2K}{\mu\eta^2}$, $c_2 = \frac{24C^2}{\mu\eta^3} + S$ and $c_3 = \frac{3072^2 K}{\mu^2\eta^4} \|L\|_\infty + \frac{\mu\eta^2}{2} + C_S$. With the presampling stage and Lemma 1, we can bound $\|L\|_\infty$ by

$$\|L\|_\infty \leq \frac{\sum_j \sigma_j^2 \text{Cof}(\Gamma)_{jj}}{\sigma_{\min} \sqrt{\text{Cof}(\Gamma)}_{\min}} \left(\sum_j \sigma_j \sqrt{\text{Cof}(\Gamma)_{jj}} \right).$$

We conclude the proof using the fact that $R(T) = \frac{1}{T}(L(p_T) - L(p^\star))$. □

F. Analysis of the case $K > d$

F.1. Proof of Theorem 3

Proof. In order to ensure that L is smooth we pre-sample each covariate n times. We note $\alpha = n/T \in (0, 1)$. This forces p_i to be greater than α for all i. Therefore L is C_S-smooth with $C_S \leq \frac{2 \max_k \text{Cof}(\Gamma)_{kk} \sigma_{\max}^2}{\alpha^3 \det(\Gamma)} \doteq \frac{C}{\alpha^3}$.

We use a similar analysis to the one of (Berthet & Perchet, 2017). Let us note $\rho_t \doteq L(p_t) - L(p^\star)$ and $\varepsilon_{t+1} \doteq (e_{\pi(t+1)} - e_{\star_{t+1}})^\top \nabla L(p_t)$ with $e_{\star_{t+1}} = \arg\max_{p \in \Delta^K} p^\top \nabla L(p_t)$. (Berthet & Perchet, 2017, Lemma 12) gives for $t \geq nK$,

$$(t+1)\rho_{t+1} \leq t\rho_t + \varepsilon_{t+1} + \frac{C_S}{t+1}.$$

Summing for $t \geq nK$ gives

$$T\rho_T \leq nK\rho_{nK} + C_S \log(eT) + \sum_{t=nK}^{T} \varepsilon_t$$

$$L(p_T) - L(p^\star) \leq K\alpha(L(p_{nK}) - L(p^\star)) + \frac{C}{\alpha^3} \frac{\log(eT)}{T} + \frac{1}{T} \sum_{t=nK}^{T} \varepsilon_t.$$

We bound $\sum_{t=nK}^{T} \varepsilon_t / T$ as in Theorem 3 of (Berthet & Perchet, 2017) by $4\sqrt{\frac{3K \log(T)}{T}} + \left(\frac{\pi^2}{6} + K \right) \frac{2\|\nabla L\|_\infty + \|L\|_\infty}{T} = \mathcal{O}\left(\sqrt{\frac{\log(T)}{T}} \right)$.

We are now interested in bounding $\alpha(L(p_{nK}) - L(p^\star))$.

By convexity of L we have

$$L(p_{nK}) - L(p^\star) \leq \langle \nabla L(p_{nK}), p_{nK} - p^\star \rangle \leq \|\nabla L(p_{nK})\|_2 \|p_{nK} - p^\star\|_2 \leq 2 \|\nabla L(p_{nK})\|_2 .$$

We have also

$$\frac{\partial L}{\partial p_k}(p_{nK}) = - \left\| \Omega(p_{nK})^{-1} \frac{X_k}{\sigma_k} \right\|_2^2 .$$

Proposition S1 shows that

$$\left\|\Omega(p)^{-1}\right\|_2 \leq \frac{1}{\lambda_{\min}(\Gamma)} \frac{\sigma_{\max}^2}{\min_k p_k} .$$

In our case, $\min_k p_{nK} = 1/K$. Therefore

$$\left\|\Omega(p_{nK})^{-1}\right\|_2 \leq \frac{K\sigma_{\max}^2}{\lambda_{\min}(\Gamma)} .$$

And finally we have

$$\|\nabla L(p_{nK})\|_2 \leq \frac{K}{\sqrt{\lambda_{\min}(\Gamma)}} \frac{\sigma_{\max}}{\sigma_{\min}} .$$

We note $C_1 \doteq \frac{2K^2}{\sqrt{\lambda_{\min}(\Gamma)}} \frac{\sigma_{\max}}{\sigma_{\min}}$. This gives

$$L(p_T) - L(p^\star) \leq \alpha C_1 + \frac{C}{\alpha^3} \frac{\log(T)}{T} + \mathcal{O}\left(\sqrt{\frac{\log(T)}{T}} \right) .$$

The choice of $\alpha = T^{-1/4}$ finally gives

$$L(p_T) - L(p^\star) = \mathcal{O}\left(\frac{\log(T)}{T^{1/4}} \right) .$$

□

F.2. Proof of Theorem 4

Proof. For simplicity we consider the case where $d = 1$ and $K = 2$. Let us suppose that there are two points X_1 and X_2 that can be sampled, with variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1 + \Delta > 1$, where $\Delta \leq 1$. We suppose also that $X_1 = X_2 = 1$ such that both points are identical.

The loss function associated to this setting is

$$L(p) = \left(\frac{p_1}{\sigma_1^2} + \frac{p_2}{\sigma_2^2} \right)^{-1} = \frac{1 + \Delta}{p_2 + p_1(1 + \Delta)} = \frac{1 + \Delta}{1 + \Delta p_1}.$$

The optimal p has all the weight on the first covariate (of lower variance): $p^\star = (1, 0)$ and $L(p^\star) = 1$.

Therefore

$$L(p) - L(p^\star) = \frac{1 + \Delta}{1 + \Delta p_1} - 1 = \frac{p_2 \Delta}{1 + \Delta p_1} \geq \frac{\Delta}{2} p_2 .$$

We see that we are now facing a classical 2-arm bandit problem: we have to choose between arm 1 giving expected reward 0 and arm 2 giving expected reward $\Delta/2$. Lower bounds on multi-armed bandits problems show that

$$\mathbb{E}L(p_T) - L(p^\star) \gtrsim \frac{1}{\sqrt{T}} .$$

Thus we obtain

$$R(T) \gtrsim \frac{1}{T^{3/2}} .$$

□

G. Geometric Interpretation

G.1. Proof of Proposition 4

Proof. We want to minimize L on the simplex Δ^K. Let us introduce the Lagrangian function

$$\mathcal{L} : (p_1, \dots, p_K, \lambda, \mu_1, \dots, \mu_K) \in \mathbb{R}^K \times \mathbb{R} \times \mathbb{R}_+^K \mapsto L(p) + \lambda \left(\sum_{k=1}^K p_k - 1 \right) - \langle \mu, p \rangle$$

Applying Karush-Kuhn-Tucker theorem gives that $p^\star$ verifies

$$\forall k \in [d], \ \frac{\partial \mathcal{L}}{\partial p_k}(p^\star) = 0.$$

Consequently

$$\forall k \in [d], \ \left\| \Omega(p^\star)^{-1} \frac{X_k}{\sigma_k} \right\|_2^2 = \lambda - \mu_k \leq \lambda.$$

This shows that the points X_k/σ_k lie within the ellipsoid defined by the equation $x^\top \Omega(p^\star)^{-2} x \leq \lambda$.

□

G.2. Geometric illustrations

In this section we present figures detailing the geometric interpretation discussed in Section 5.

Geometrically the dual problem (D) is equivalent to finding an ellipsoid containing all data points X_k/σ_k such that the sum of the inverse of the semi-axis is maximized. The points that lie on the boundary of the ellipsoid are the one that have to be sampled. We see here that we have to sample the points that are far from the origin (after being rescaled by their standard deviation) because they cause less uncertainty.

We see that several cases can occur as shown on Figure 1. If one covariate is in the interior of the ellipsoid it is not sampled because of the KKT equations (see Proposition 4). However if all the points are on the ellipsoids some of them may not be sampled. It is the case on Figure 1(b) where X_1 is not sampled. This is due to the fact that a little perturbation of another point, for example X_3 can change the ellipsoid such that X_1 ends up inside the ellipsoid as shown on Figure 1(d). This case can consequently be seen as a limit case.

References

Berthet, Q. and Perchet, V. Fast rates for bandit optimization with upper-confidence frank-wolfe. In *Advances in Neural Information Processing Systems*, pp. 2225–2234, 2017.

Carpentier, A., Lazaric, A., Ghavamzadeh, M., Munos, R., and Auer, P. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pp. 189–203. Springer, 2011.

Chafaï, D., Guédon, O., Lecué, G., and Pajor, A. *Interactions between compressed sensing random matrices and high dimensional geometry*. Citeseer, 2012.

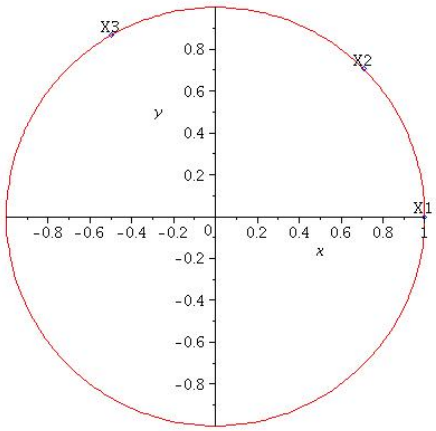
Maurer, A. and Pontil, M. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

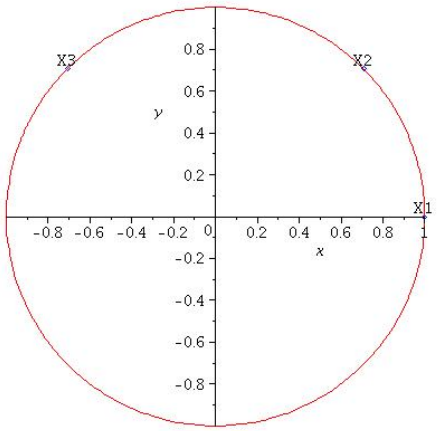
Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, Q. and Chen, W. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Neural Information Processing Systems*, mar 2017. URL http://arxiv.org/abs/1703.01610.

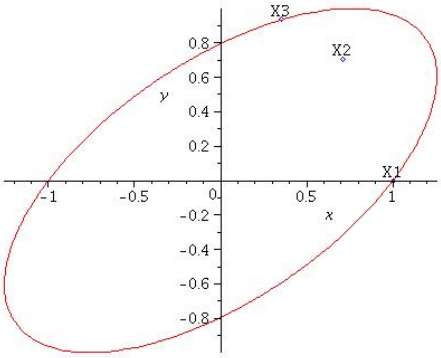
Whittle, P. A multivariate generalization of tchebichev's inequality. *The Quarterly Journal of Mathematics*, 9(1):232–240, 1958.



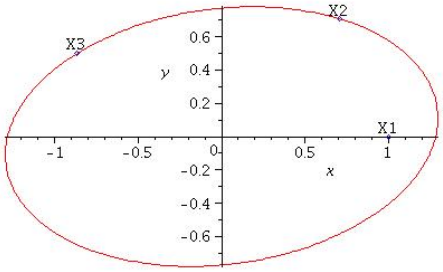
(a) $p_1 = 0.21 \quad p_2 = 0.37 \quad p_3 = 0.42$



(b) $p_1 = 0 \quad p_2 = 0.5 \quad p_3 = 0.5$



(c) $p_1 = 0.5 \quad p_2 = 0 \quad p_3 = 0.5$



(d) $p_1 = 0 \quad p_2 = 0.5 \quad p_3 = 0.5$

Figure 1. Different minimal ellipsoids