A Proofs for Stochastic Exp-Concave Optimization

Proof (of Lemma 1) The exp-concavity of $f \mapsto \ell(f, z)$ for each $z \in \mathcal{Z}$ implies that, for all $z \in \mathcal{Z}$ and all distributions Q over $\mathcal{F}$:

$$\mathsf{E}_{f \sim Q}\left[e^{-\eta \ell(f,z)}\right] \leq e^{-\eta \ell(\mathsf{E}_{f \sim Q}[f], z)} \quad \Longleftrightarrow \quad \ell(\mathsf{E}_{f \sim Q}[f], z) \leq -\frac{1}{\eta} \log \mathsf{E}_{f \sim Q}\left[e^{-\eta \ell(f,z)}\right].$$

It therefore holds that for all distributions P over $\mathcal{Z}$, for all distributions Q over $\mathcal{F}$, there exists (from convexity of $\mathcal{F}$) $f^* = \mathsf{E}_{f \sim Q}[f] \in \mathcal{F}$ satisfying

$$\mathsf{E}_{Z \sim P}[\ell(f^*, Z)] \leq \mathsf{E}_{Z \sim P}\left[-\tfrac{1}{\eta} \log \mathsf{E}_{f \sim Q}\left[e^{-\eta \ell(f,Z)}\right]\right].$$

This condition is equivalent to *stochastic mixability* as well as the *pseudoprobability convexity (PPC) condition*, both defined by Van Erven et al. (2015). To be precise, for stochastic mixability, in Definition 4.1 of Van Erven et al. (2015), take their $\mathcal{F}_d$ and $\mathcal{F}$ both equal to our $\mathcal{F}$, their $\mathcal{P}$ equal to $\{P\}$, and $\psi(f) = f^*$; then strong stochastic mixability holds. Likewise, for the PPC condition, in Definition 3.2 of Van Erven et al. (2015) take the same settings but instead $\phi(f) = f^*$; then the strong PPC condition holds. Now, Theorem 3.10 of Van Erven et al. (2015) states that the PPC condition implies the (strong) central condition. ■

Proof (of Theorem 1) First, from Lemma 1, the convexity of $\mathcal{F}$ together with η-exp-concavity implies that $(P, \ell, \mathcal{F})$ satisfies the η-central condition.

The remainder of the proof is a drastic simplification of the proof of Theorem 7 of Mehta and Williamson (2014). Technically, Theorem 7 of that work applies directly, but one can get substantially smaller constants by avoiding much of the technical machinery needed there to handle VC-type classes (e.g. symmetrization, chaining, Talagrand's inequality).

Denote by $\mathcal{L}_f := \ell_f - \ell_{f^*}$ the excess loss with respect to comparator f^*. Our goal is to show that, with high probability, ERM does not select any function $f \in \mathcal{F}$ whose excess risk $\mathsf{E}[\mathcal{L}_f]$ is larger than $\frac{a}{n}$ for some constant a. Clearly, with probability 1 ERM will never select any function for which both $\mathcal{L}_f \geq 0$ almost surely and with some positive probability $\mathcal{L}_f > 0$; we call these functions the empirically inadmissible functions. For any $\gamma_n > 0$, let $\mathcal{F}_{\succeq \gamma_n}$ be the subclass formed by starting with $\mathcal{F}$, retaining only functions whose excess risk is at least γ_n, and further removing the empirically inadmissible functions.

Our goal now may be expressed equivalently as showing that, with high probability, ERM does not select any function $f \in \mathcal{F}_{\succeq \gamma_n}$ where $\gamma_n = \frac{a}{n}$ and $a > 1$ is some constant to be determined later. Let $\mathcal{F}_{\succeq \gamma_n, \varepsilon}$ be an optimal proper (ε/L)-cover for $\mathcal{F}_{\succeq \gamma_n}$ in the ℓ_2 norm. From the Lipschitz property of the loss it follows that this cover induces an ε-cover in sup norm over the loss-composed function class $\{\ell_f : f \in \mathcal{F}_{\succeq \gamma_n}\}$. Observe that an ε-cover of $\mathcal{F}_{\succeq \gamma_n}$ in the ℓ_2 norm has cardinality at most $(4R/\varepsilon)^d$ (Carl and Stephani, 1990, equation 1.1.10), and the cardinality of an optimal *proper* ε-cover is at most the cardinality of an optimal $(\varepsilon/2)$-cover. (Vidyasagar, 2002, Lemma 2.1). It hence follows that $|\mathcal{F}_{\succeq \gamma_n, \varepsilon}| \leq \left(\frac{8LR}{\varepsilon}\right)^d$.

Let us consider some fixed $f \in \mathcal{F}_{\succeq \gamma_n, \varepsilon}$. As we removed the empirically inadmissible functions, there exists some $\eta_f \geq \eta$ for which $\mathsf{E}[e^{-\eta_f \mathcal{L}_f}] = 1$. Theorem 3 and Lemma 4, both from Mehta and Williamson (2014), imply that

$$\log \mathsf{E}_{Z \sim P}\left[e^{-(\eta_f/2)\mathcal{L}_f}\right] \leq -\frac{0.18 \eta_f a}{(B\eta_f \vee 1)n}.$$

Applying Theorem 1 of Mehta and Williamson (2014) with $t = \frac{a}{2n}$ and the η in that theorem set to $\eta_f/2$ yields:

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_f(Z_j) \leq \frac{a}{2n}\right) \leq \exp\left(-0.18 \frac{\eta_f}{B\eta_f \vee 1} a + \frac{a\eta_f}{4n}\right).$$

Now, since $\eta \leq \eta_f$ for all $f \in \mathcal{F}_{\succeq \gamma_n, \varepsilon}$, taking the union bound over $\mathcal{F}_{\succeq \gamma_n, \varepsilon}$ implies that

$$\Pr\left(\exists f \in \mathcal{F}_{\succeq \gamma_n, \varepsilon} : \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_f(Z_j) \leq \frac{a}{2n}\right) \leq \left(\frac{8LR}{\varepsilon}\right)^d \exp\left(-0.18 \frac{\eta}{B\eta \vee 1} a + \frac{a\eta}{4n}\right).$$

Setting $\varepsilon = \frac{1}{2n}$ and taking $n \geq 5$, from inversion it follows that with probability at least $1 - \delta$, for all $f \in \mathcal{F}_{\succeq \gamma_n, \varepsilon}$, we have $\frac{1}{n} \sum_{j=1}^n \mathcal{L}_f(Z_j) \leq \frac{a}{2n}$, where

$$a = 8 \left(B \vee \frac{1}{\eta} \right) \left(d \log(16 L R n) + \log \frac{1}{\delta} \right).$$

Now, since $\sup_{f \in \mathcal{F}_{\succeq \gamma_n}} \min_{f_\varepsilon \in \mathcal{F}_{\succeq \gamma_n, \varepsilon}} \| \ell_f - \ell_{f_\varepsilon} \|_\infty \leq \frac{1}{2n}$, and increasing a by 1 to guarantee that $a > 1$, with probability at least $1 - \delta$, for all $f \in \mathcal{F}_{\succeq \gamma_n}$, we have $\frac{1}{n} \sum_{j=1}^n \mathcal{L}_f(Z_j) > 0$. ■

Proof (of Lemma 2) The main tool we use is part 2 of Theorem 5.4 of Van Erven et al. (2015). First, as per the proof of Lemma 1, note that the central condition as defined in the present work is equivalent to the strong PPC condition of Van Erven et al. (2015). We actually can improve that result due to our easier setting because we may take their function v to be the constant function identically equal to η. Consequently, in equation (70) of Van Erven et al. (2015), we may take $\varepsilon = 0$, improving their constant c_2 by a factor of 3; moreover, their result actually holds for the second moment, not just the variance, yielding:

$$\mathsf{E}[X^2] \leq \frac{2}{\eta \kappa(-2\eta B)} \mathsf{E}[X], \tag{6}$$

where $\kappa(x) = \frac{e^x - x - 1}{x^2}$.

We now study the function

$$x \mapsto \frac{1}{\kappa(-x)} = \frac{x^2}{e^{-x} + x - 1}.$$

We claim that for all $x \geq 0$:

$$\frac{x^2}{e^{-x} + x - 1} \leq 2 + x.$$

L'Hôpital's rule implies that the inequality holds for $x = 0$, and so it remains to consider the case of $x > 0$.

First, observe that the denominator is nonnegative, and so we may rewrite this inequality as

$$x^2 \leq (2 + x)(e^{-x} + x - 1),$$

which simplifies to

$$0 \leq 2e^{-x} + x + xe^{-x} - 2 \qquad \Leftrightarrow \qquad 2(1 - e^{-x}) \leq x(1 + e^{-x}).$$

Therefore, we just need to show that, for all $x > 0$,

$$\frac{2}{x} \leq \frac{1 + e^{-x}}{1 - e^{-x}} = \frac{e^{x/2} + e^{-x/2}}{e^{x/2} - e^{-x/2}} = \coth(x/2),$$

which is equivalent to showing that for all $x > 0$,

$$\tanh(x) \leq x.$$

But this indeed holds, since

$$\begin{aligned} \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} &= \frac{2(x + \frac{x^3}{3!} + \frac{x^5}{5!} + \dots)}{2(1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \dots)} \\ &= x \cdot \frac{1 + \frac{x^2}{3!} + \frac{x^4}{5!} + \dots}{1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \dots} \\ &\leq x. \end{aligned}$$

The desired inequality is now established.

Returning to (6), we have

$$\mathsf{E}[X^2] \leq \frac{2}{\eta}(2 + 2\eta B) \mathsf{E}[X] \leq 4 \left(\frac{1}{\eta} + B \right) \mathsf{E}[X].$$

■

Proof (of Lemma 3) The following simple version of Bernstein's inequality will suffice for our analysis. Let $X_1, \ldots, X_n$ be independent random variables satisfying $X_j \geq B$ almost surely. Then

$$\Pr\left(\frac{1}{n}\sum_{j=1}^{n} X_j - \mathsf{E}[X] \geq t\right) \leq \exp\left(-\frac{nt^2}{2\left(\mathsf{E}\left[\frac{1}{n}\sum_{j=1}^{n} X^2\right] + \frac{Bt}{3}\right)}\right).$$

Denote by $\mathcal{L}_f := \ell_f - \ell_{g_1}$ the excess loss with respect to comparator g_1. Fix some $f \in \mathcal{G}' \setminus \{g_1\}$, take $X = -\mathcal{L}_f$, and set $t = \mathsf{E}[\mathcal{L}_f]$, yielding:

$$\begin{aligned}
\Pr\left(\frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_f(Z_j) \leq 0\right) &\leq \exp\left(-\frac{n\,\mathsf{E}[\mathcal{L}_f]^2}{2(\mathsf{E}[\mathcal{L}_f^2] + \frac{1}{3}B\,\mathsf{E}[\mathcal{L}_f])}\right) \\
&\leq \exp\left(-\frac{n\,\mathsf{E}[\mathcal{L}_f]^2}{2(C\,\mathsf{E}[\mathcal{L}_f]^q + \frac{1}{3}B\,\mathsf{E}[\mathcal{L}_f])}\right) \\
&= \exp\left(-\frac{n\,\mathsf{E}[\mathcal{L}_f]^{2-q}}{2\left(C + \frac{1}{3}B\,\mathsf{E}[\mathcal{L}_f]^{1-q}\right)}\right) \\
&\leq \exp\left(-\frac{n\,\mathsf{E}[\mathcal{L}_f]^{2-q}}{2\left(C + \frac{1}{3}B^{2-q}\right)}\right).
\end{aligned}$$

Therefore, if

$$\mathsf{E}[\mathcal{L}_f] \geq \left(\frac{2\left(C + \frac{B^{2-q}}{3}\right)\log\frac{|\mathcal{G}'|}{\delta}}{n}\right)^{1/(2-q)}, \tag{7}$$

then it holds with probability at least $1 - \frac{\delta}{|\mathcal{G}'|-1}$ that $\frac{1}{n}\sum_{j=1}^{n} \mathcal{L}_f(Z_j) > 0$. The result follows by taking a union bound over the subclass of $\mathcal{G}' \setminus \{g_1\}$ for which (7) holds. ■

B Proofs for Model Selection Aggregation (Section 7)

Proof (of Theorems 4 and 5) The starting point is the following bound for the progressive mixture rule when run with prior π and parameter η, due to Audibert (see Theorem 1 of Audibert (2008), but the result was already proved in an earlier technical report version of Audibert (2009) (see Corollary 4.1 and Lemma 3.3 therein). When run on an n-sample, an online-to-batch conversion of the progressive mixture rule yields a hypothesis $\hat{f}$ satisfying

$$\mathsf{E}_{Z^n}\left[\mathsf{E}_Z\left[\ell(Y, \hat{f}(X))\right]\right] \leq \inf_{\rho \in \Delta(\mathcal{F})}\left\{\mathsf{E}_{f\sim\rho}\,\mathsf{E}_Z\left[\ell(Y, f(X))\right] + \frac{D(\rho \,\|\, \pi)}{\eta(n+1)}\right\}$$

where $D(\rho \,\|\, \pi)$ is the KL-divergence of ρ from π.[3] Note that this bound does not explicitly depend on the boundedness nor the Lipschitz continuity of the loss.

Fix some ρ^* that nearly obtains the infimum (or obtains it, if possible). Then

$$\mathsf{E}_{Z^n}\left[\mathsf{E}_Z\left[\ell(Y, \hat{f}(X))\right]\right] - \mathsf{E}_{f\sim\rho^*}\,\mathsf{E}_Z\left[\ell(Y, f(X))\right] \leq \frac{D(\rho^* \,\|\, \pi)}{\eta(n+1)}.$$

We cannot apply the boosting the confidence trick just yet as the LHS is not a nonnegative random variable;

[3] We say "of ρ from π" because the Bregman divergence form of the KL-divergence, which makes clear that the KL-divergence is measure of the curvature of negative Shannon entropy between ρ and π when considering a first-order Taylor expansion around π.

this issue motivates the following rewrite.

$$\begin{aligned}
&\mathsf{E}_{Z^n}\left[\mathsf{E}_Z\left[\ell(Y,\hat{f}(X))\right]\right]-\mathsf{E}_Z\left[\ell(Y,f^*(X))\right] \\
&\leq \underbrace{\mathsf{E}_{f\sim\rho^*}\mathsf{E}_Z\left[\ell(Y,f(X))\right]-\mathsf{E}_Z\left[\ell(Y,f^*(X))\right]}_{\textsc{gap}(\rho^*,f^*)}+\frac{D(\rho^*\,\|\,\pi)}{\eta(n+1)}.
\end{aligned}$$

When the progressive mixture rule is run on K independent samples, yielding hypotheses $f^{(1)},\ldots,f^{(K)}$, then Markov's inequality implies that with probability at least $1-e^{-K}$ (over the (Kn)-sample) there exists $j\in[K]$ for which

$$\begin{aligned}
&\mathsf{E}_Z\left[\ell(Y,f^{(j)}(X))\right]-\mathsf{E}_Z\left[\ell(Y,f^*(X))\right] \\
&\leq e\left(\textsc{gap}(\rho^*,f^*)+\frac{D(\rho^*\,\|\,\pi)}{\eta(n+1)}\right),
\end{aligned}$$

which can be re-expressed as

$$\begin{aligned}
&\mathsf{E}_Z\left[\ell(Y,f^{(j)}(X))\right]-\mathsf{E}_Z\left[\ell(Y,f^*(X))\right] \\
&\leq e\cdot\textsc{gap}(\rho^*,f^*)+\frac{e\cdot D(\rho^*\,\|\,\pi)}{\eta(n+1)} \\
&= e\left(\inf_{\rho\in\Delta(\mathcal{F})}\left\{\mathsf{E}_{f\sim\rho}\mathsf{E}_Z\left[\ell(Y,f(X))\right]+\frac{D(\rho\,\|\,\pi)}{\eta(n+1)}\right\}-\mathsf{E}_Z\left[\ell(Y,f^*(X))\right]\right) \\
&= e\cdot\textsc{BayesRed}_\eta\left(n,\pi\right).
\end{aligned}$$

In the sequel, we assume that this high probability event has occurred.

Now, let $\widetilde{\mathcal{F}}=\operatorname{conv}\left(\{f^{(1)},\ldots,f^{(K)}\}\right)$. Clearly, $f^{(j)}\in\widetilde{\mathcal{F}}$, and so we also have

$$\inf_{f\in\widetilde{\mathcal{F}}}\mathsf{E}_Z\left[\ell(Y,f(X))\right]\leq e\cdot\textsc{BayesRed}_\eta\left(n,\pi\right). \tag{8}$$

It therefore is sufficient to learn over $\widetilde{\mathcal{F}}$ and compete with its risk minimizer. But this is only a K-dimensional problem, and if $\delta=e^{-K}$, we have $K=\log\frac{1}{\delta}$. To see why the problem is only K-dimensional, consider the transformed problem, where

$$\tilde{x}=\begin{pmatrix}f^{(1)}(x)\\ \vdots\\ f^{(K)}(x)\end{pmatrix}.$$

The loss can now be reparameterized, from

$$\ell\colon\widetilde{\mathcal{F}}\to\mathbb{R}\quad\text{with}\quad\ell\colon f\mapsto\ell(y,f(x))$$

to

$$\tilde{\ell}\colon\Delta^{K-1}\to\mathbb{R}\quad\text{with}\quad\tilde{\ell}\colon q\mapsto\ell(y,\langle q,\tilde{x}\rangle),$$

where Δ^{K-1} is the $(K-1)$-dimensional simplex $\left\{q\in[0,1]^K\colon\sum_{j=1}^K q_j=1\right\}$.

Δ^{K-1} is clearly convex and the loss is η-exp-concave with respect to $q\in\Delta^{K-1}$; to see the latter, observe that from the η-exp-concavity of the loss with respect to $\hat{y}=\langle q,\tilde{x}\rangle$:

$$\begin{aligned}
\mathsf{E}_{q\sim P_q}\left[e^{-\eta\ell(y,\langle q,\tilde{x}\rangle)}\right]&\leq e^{-\eta\ell(y,\mathsf{E}_{q\sim P_q}[\langle q,\tilde{x}\rangle])} \\
&= e^{-\eta\ell(y,\langle\mathsf{E}_{q\sim P_q}[q],\tilde{x}\rangle)}.
\end{aligned}$$

Lastly, the loss is still bounded by B since $\widetilde{\mathcal{F}}$ consists only of convex aggregates of $\hat{f}_1, \ldots, \hat{f}_K$, themselves convex aggregates over $\mathcal{F}$ (and we assumed boundedness of the loss with respect to the original class).

We now can proceed in two ways. The high probability bound for EWOO (the first display after Corollary 2) applies immediately. This bound can be simplified to (taking $d = K = \lceil \log(2/\delta) \rceil$)

$$O\left(\frac{\sqrt{B} \left(\log \frac{1}{\delta} + \sqrt{\log \frac{1}{\delta}} \log n \right)}{\eta n} + \frac{B \left(\log \log n + \log \frac{1}{\delta} \right)}{n} \right),$$

which, in light of (8), proves Theorem 4.

If we further assume the loss framework of Gonen and Shalev-Shwartz (2016), then $\tilde{\ell}$ still satisfies α-strong convexity in the sense needed because, conditional on the actual prediction $\hat{y}$, the loss $\tilde{\ell}$ is the same as loss ℓ. Hence, the bound (5) ConfidenceBoost from Corollary 1 applies (taking $d = K = \lceil \log \frac{3}{\delta} \rceil$), finishing the proof of Theorem 5. ■