Fast Rates with High Probability in Exp-concave Statistical Learning

Nishant A. Mehta

Exp-concave Statistical Learning

Goal: Obtain low Excess Risk $\underbrace{\mathrm{E}_P\, \ell_{\hat{f}}(Z)}_{R(\hat{f})} - \inf_{f \in \mathcal{F}} \underbrace{\mathrm{E}_P\, \ell_f(Z)}_{R(f)}$

Stochastic Exp-concave Optimization (SECO)

$\mathcal{F}$ is convex subset of $\mathbb{R}^d$

$f \mapsto \ell_f(z)$ is exp-concave for all $z \in \mathcal{Z}$

$\sup_{z \in \mathcal{Z}} \sup_{f,g \in \mathcal{F}} |\ell_f(z) - \ell_g(z)| \le B$ (bounded loss)

$\|\nabla_f \ell(f,z) - \nabla_f \ell(f',z)\| \le \beta \|f - f'\|_2$ (β-smooth loss)

$\sup_{f \in \mathcal{F}} \|f\|_2^2 - \inf_{f \in \mathcal{F}} \|f\|_2^2 \le 2D^2$ (bounded diameter)

Model Selection Aggregation (with exp-concave loss)

$\mathcal{F}$ is countable

$Z = (X, Y)$ and $\ell_f(Z) = \ell(Y, f(X))$

$\sup_{f,f' \in \mathcal{F}} |\ell(y, f(x)) - \ell(y, f'(x))| \le B$ for all (x, y) (bounded loss)

$\hat{y} \mapsto \ell(y, \hat{y})$ is η-exp-concave and L-Lipschitz

Previous Results

SECO

In Expectation guarantees on excess risk

Penalized ERM (PERM): $\psi(n) = O\left(\frac{1}{n}\left(\frac{\beta d}{\eta} + Bd\right) + D^2\right)$
[Koren and Levy, 2015]

Exponentially Weighted Online Optimization (EWOO): $\frac{d(1 + \log(n+1))}{\eta n}$
[Hazan, Agarwal, Kale, 2007]

High probability ($1 - \delta$) guarantees on excess risk

ERM: $\tilde{O}\left(\frac{(LD \vee \frac{1}{\eta})\left(d \log n + \log \frac{1}{\delta}\right) + 1}{n}\right)$

Online Newton Step: $\tilde{O}\left(\frac{d(LD \vee \frac{1}{\eta})\left(\log n + \log \frac{1}{\delta}\right)}{n}\right)$
[Hazan, Agarwal, Kale, 2007]

Open: Excess risk rate of $\frac{d}{\eta n}$ with logarithmic dependence on δ

Model Selection Aggregation (High probability guarantees)

Empirical Star algorithm: $O\left(\frac{\log|\mathcal{F}| + \log \frac{1}{\delta}}{n}\right)$
[Audibert, 2008]

Q-aggregation: $R(\hat{f}) \le \min_{f \in \mathcal{F}} \left\{ R(f) + O\left(\frac{1}{n} \log \frac{1}{\pi(f)}\right)\right\} + O\left(\frac{\log \frac{1}{\delta}}{n}\right)$

[Lecué and Rigollet, 2014]

Main Result (SECO)

ConfidenceBoost Algorithm

Given: $\mathbf{Z}_1, \ldots, \mathbf{Z}_K \overset{\text{iid}}{\sim} P^{n_{\text{I}}}$, $\mathbf{Z}_{\text{II}} \sim P^{n_{\text{II}}}$, and learner $\mathcal{A}_{\mathcal{F}}$
1) For $j = 1, \ldots, K$
$\hat{f}_j = \mathcal{A}_{\mathcal{F}}(\mathbf{Z}_j)$
2) Return empirical risk minimizer over $\{\hat{f}_1, \ldots, \hat{f}_K\}$ on sample $\mathbf{Z}_{\text{II}}$

Theorem
With probability at least $1 - \delta$, ConfidenceBoost run with PERM, $K = \lceil \log(2/\delta) \rceil$, $n_{\text{I}} = n/(2K)$, and $n_{\text{II}} = n/2$ obtains excess risk

$$O\left(\frac{\log \frac{1}{\delta}}{n}\left(\frac{d\beta}{\eta} + dB + D\right)\right)$$

Analysis I: Exp-concavity to Bernstein

Exp-concavity

$$e^{-\eta \ell_{\mathrm{E}_{f \sim \Pi}[f]}(z)} \ge \mathrm{E}_{f \sim \Pi}\left[e^{-\eta \ell_f(z)}\right] \quad \text{for all } \Pi \in \Delta(\mathcal{F})$$

$$\left(\text{so } \ell_{\mathrm{E}_{f \sim \Pi}[f]}(z) \le -\frac{1}{\eta} \log \mathrm{E}_{f \sim \Pi}\left[e^{-\eta \ell_f(z)}\right]\right)$$

$$\Downarrow$$

Central Condition

$$\mathrm{E}_{Z \sim P}\left[\ell_{f^*}(Z)\right] \le \mathrm{E}_{Z \sim P}\left[-\frac{1}{\eta} \log \mathrm{E}_{f \sim \Pi}\left[e^{-\eta \ell(Z)}\right]\right] \quad \text{for all } \Pi \in \Delta(\mathcal{F})$$

If $\sup_{f \in \mathcal{F}} \sup_{z \in \mathcal{Z}} |\ell_f(z) - \ell_{f^*}(z)| \le B$ (loss is bounded by B)

and the η-central condition holds

Then a **Bernstein Condition** holds:

$$\mathrm{E}\left[(\ell_f - \ell_{f^*})^2\right] \le 4\left(\frac{1}{\eta} + B\right) \mathrm{E}\left[\ell_f - \ell_{f^*}\right] \text{ for all } f \in \mathcal{F}$$

Analysis II: Boosting the Confidence

Assume that problem $(P, \ell, \mathcal{F})$ satisfies a C-Bernstein condition. Let $\mathcal{F}_K = \{f_1, \ldots, f_K\}$ be a subclass with risk minimizer f_1

Then an approximate Bernstein condition holds for $(P, \ell, \mathcal{F}_K)$:

$$\mathrm{E}\left[(\ell_{f_j} - \ell_{f_1})^2\right] \le C\left(3\,\mathrm{E}[\ell_{f_j} - \ell_{f_1}] + 4\,\mathrm{E}[\ell_{f_1} - \ell_{f^*}]\right) \text{ for all } j \in [K]$$

Run Penalized ERM K times, each on an independent n-sample, yielding hypotheses $\{\hat{f}_1, \ldots, \hat{f}_K\}$. With probability at least $1 - e^{-K}$, there is a $j^* \in [K]$ for which $\hat{f}_{j^*}$ has excess risk at most $e \cdot \psi(n)$.

Thus, with probability at least $1 - e^{-K}$, a Bernstein condition holds for a subclass of $\mathcal{F}_K$ where we include only $\hat{f}_{j^*}$ and functions with excess risk at least $e \cdot \psi(n)$.

Hence, ERM on the subclass achieves low excess risk over $\hat{f}_{j^*}$.

High Probability Bounds for Online-to-Batch Conversions of Online Exp-concave Learners

(Supervised) online convex optimization

Adversary selects sequence of n points $z_1, \ldots, z_n$

In round t:
1) Learner plays prediction f_t
2) Adversary reveals z_t
3) Learner suffers loss $\ell_{f_t}(z_t)$

Goal: Obtain low Regret $\sum_{t=1}^n \ell_{f_t}(z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \ell_f(z_t)$

Now, suppose adversary draws $z_1, \ldots, z_n$ iid from distribution P. Given a regret bound $\mathcal{R}_n$, we want an excess risk bound at the same level.

Define the martingale difference sequence

$$\xi_t = \mathrm{E}_Z\left[\ell_{f_t}(Z) - \ell_{f^*}(Z)\right] - \left(\ell_{f_t}(Z_t) - \ell_{f^*}(Z_t)\right)$$

Suppose $\mathrm{Var}[\xi_t \mid Z_t^{t-1}] \le C\, \mathrm{E}_Z[\ell_{f_t}(Z) - \ell_{f^*}(Z)]$

Then online-to-batch conversion $\bar{f}_n := \frac{1}{n}\sum_{t=1}^n f_t$ obtains excess risk at the rate $\frac{1}{n}\mathcal{R}_n$ with high probability.

[Kakade and Tewari, 2009] - They considered Lipschitz, strongly convex functions. Key idea: Freedman's inequality for martingales

It turns out that under exp-concavity and bounded loss:

$$\mathrm{Var}[\xi_t \mid Z_t^{t-1}] \le 4\left(\frac{1}{\eta} + B\right) \mathrm{E}_Z[\ell_{f_t}(Z) - \ell_{f^*}(Z)]$$

Theorem
If loss is η-exp-concave and B-bounded and an online learner has regret bound $\mathcal{R}_n$, then with probability at least $1 - \delta$, an online-to-batch conversion of the learner obtains excess risk

$$O\left(\frac{1}{n}\left(\mathcal{R}_n + \left(\frac{1}{\eta} + B\right) \log \frac{\log n}{\delta}\right)\right)$$

Corollary
Taking the online exp-concave learner from the previous result to be EWOO and plugging in the corresponding regret bound yields

$$O\left(\frac{d \log n}{\eta n} + \frac{B \log \frac{\log n}{\delta}}{n} + \frac{\sqrt{d \log n}}{n}\left(\sqrt{\frac{B \log \log n}{\eta}} + \sqrt{\left(\frac{1}{\eta^2} + \frac{B}{\eta}\right) \log \frac{1}{\delta}}\right)\right)$$

Model Selection Aggregation

Progressive Mixture Rule Algorithm (Not new!)

1) Start with a prior $\pi \in \Delta(\mathcal{F})$
2) For rounds $t = 1, \ldots, n$
Play $\hat{f}_t = \mathrm{E}_{f \sim \rho_t}[f]$, where $\rho_t(f) \propto e^{-\eta \sum_{j=0}^{t-1} \ell(Y_j, f(X_j))} \cdot \pi(f)$
3) Return $\frac{1}{n}\sum_{t=1}^n \hat{f}_t$

Progressive mixture rule satisfies in-expectation risk guarantee: [Audibert, 2008]

$$\mathrm{E}_{Z^n}\left[R(\hat{f})\right] \le \underbrace{\inf_{\rho \in \Delta(\mathcal{F})}\left\{\mathrm{E}_{f \sim \rho}[R(f)] + \frac{\mathrm{D}_{\mathrm{KL}}(\rho \,\|\, \pi)}{\eta n}\right\}}_{\text{Bayesian Redundancy: } \mathrm{BayesRed}_\eta(n, \pi)}$$

$$\mathrm{E}_{Z^n}\left[\underbrace{R(\hat{f}) - R(f^*)}_{\text{nonnegative r.v.}}\right] \le \inf_{\rho \in \Delta(\mathcal{F})}\left\{\underbrace{\mathrm{E}_{f \sim \rho}[R(f)] - R(f^*)}_{\mathrm{GAP}(\rho, f^*)} + \frac{\mathrm{D}_{\mathrm{KL}}(\rho \,\|\, \pi)}{\eta n}\right\}$$

$\downarrow$ boosting the confidence trick

K hypotheses $f^{(1)}, \ldots, f^{(K)}$

With probability at least $1 - e^{-K}$, there is a $j \in [K]$ such that:

$$R(f^{(j)}) - R(f^*) \le e \cdot \inf_{\rho \in \Delta(\mathcal{F})}\left\{\mathrm{E}_{f \sim \rho}[R(f)] - R(f^*) + \frac{\mathrm{D}_{\mathrm{KL}}(\rho \,\|\, \pi)}{\eta n}\right\}$$

Thus, sufficient to compete with the class $\tilde{\mathcal{F}} = \mathrm{conv}(f^{(1)}, \ldots, f^{(K)})$

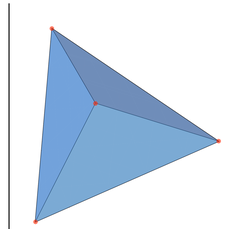
Reduction to Stochastic Exp-concave Optimization

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

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

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

EWOO on convex hull



Theorem
With probability at least $1 - \delta$, EWOO run on the convex hull of K calls to the Progressive Mixture Rule obtains excess risk at most

$$\mathrm{BayesRed}_\eta\left(\frac{n}{2\lceil \log \frac{2}{\delta} \rceil}, \pi\right) + O\left(\frac{\sqrt{B}\left(\log \frac{1}{\delta} + \sqrt{\log \frac{1}{\delta} \log n}\right)}{\eta n} + \frac{B \log \frac{\log n}{\delta}}{n}\right)$$

- First high probability bound that improves with prior measure on subclass of nearly optimal hypotheses
- No dependence on Lipschitz constant of the loss!