LOG CONCAVE COUPLING FOR SAMPLING FROM NEURAL NET POSTERIOR DISTRIBUTIONS

ANDREW R. BARRON, CURTIS MCDONALD

Classification AMS 2020: 60-08, 62-08, 62G08, 62M45, 68Q25, 68Q32, 68Q87, 68T07

Keywords: Neural Networks, Computational Learning Theory, Posterior Sampling, Markov Chain Monte Carlo, Log Concavity, High Dimensionality, Online Learning

Abstract: A method to rapidly sample neural net posteriors is provided. Parameters w are coupled with an auxiliary random vector ξ such that both are fast to sample. The distribution of w given ξ is log-concave, permiting rapid sampling from this conditional and rapid computation of the gradient of log density of ξ. Using this gradient as the drift allows a stochastic diffusion for sampling ξ. The density of ξ is shown to be strictly log-concave in high dimensions, so it mixes rapidly.

In addition to the posterior sampling, some clarification of the associated statistical risk is provided in this proceedings extended abstract of the work presented at [1]. It is a companion to [2],[3] which provide analogous greedy Bayes sampling results.

Function Model: Consider single hidden-layer nets $f_w(x) = f(x,w) = \sum_{k=1}^K c_k \psi(\underline{w}_k \cdot x)$ with K units, with $\sum_j |w_{j,k}| \le 1$, with activation $\psi(z)$ having $|\psi(z)| \le 1$, $|\psi'(z)| \le 1$ and $|\psi''(z)| \le 1$ on $-1 \le z \le 1$. Fix positive V and $c_k = \pm V/K$, though Bayesian models to adapt V, K and the c_k are possible. Such networks provide accurate approximation [4] for $f(x)$ with f/V in the convex hull of $\Psi = \{\pm\, \psi(w \cdot x),\, x \in [-1,1]^d : ||w||_1 \le 1\}$. A coordinate of x is -1 to allow shifts. Set the sign of c_k to be $+$ when $\psi(z)$ is odd symmetric, such as a tanh. In general, among the K terms, a number are set positive and a number negative. An example activation is $\psi(z) = \frac{1}{2} z_+^2$, a squared rectified linear unit.

Data Model: Stochastic data (X_i, Y_i), $1 \le i \le n$, are independent from a distribution $P_{X,Y}$ with marginal P_X on $[-1,1]^d$ and conditional $P_{Y|X}$ with mean $f(X)$. For statistical risk bounds the conditional is normal of specified variance σ^2, or, more generally, an arbitrary distribution with finite variance bounded by σ^2. For online learning regret bounds, the data are an arbitrary sequence of bounded inputs and outputs.

Posterior for Computation: Estimation proceeds by sampling from a posterior. The prior $p_0(w)$ on $w = (\underline{w}_1, ..., \underline{w}_K)$ makes each $\underline{w}_k$ independent uniform on the simplex $S_1^d = \{w : \sum_{j=1}^d |w_j| \le 1\}$. Let $\ell(w) = \frac{1}{2}\sum_{i=1}^n (res_i(w))^2$ with $res_i(w) = y_i - \sum_{k=1}^K c_k \psi(x_i \cdot \underline{w}_k)$. With the normal, the posterior is proportional to $p_0(w) \exp\{-\frac{1}{\sigma^2}\ell(w)\}$. More generally, we sample from $p(w) = p_n(w)$ proportional to $p_0(w)\exp\{-\beta \ell(w))\}$ with some $\beta > 0$ as a computational device, with sum of squared residuals in the exponent, even if the deviations of Y are not normal. To estimate the function using $\hat{f}(x) = \int f(x,w) p(w) dw$ we draw independent samples from $p(w)$ and average $f(x,w)$.

Non-Logconcavity of *p(w)*: Let $H(w) = \nabla \nabla^{\mathrm{T}} \ell(w)$ be the Hessian of $\ell(w)$ and consider the quadratic form $a^{\mathrm{T}} H(w) a$ where a in R^{Kd} has blocks a_k in R^d. It takes the form

$$\sum_{i=1}^n \big(\sum_{k=1}^K c_k \psi'(x_i \cdot w_k) a_k \cdot x_i\big)^2 - \sum_{i=1}^n res_i(w) \sum_{k=1}^K c_k \psi''(x_i \cdot w_k)(a_k \cdot x_i)^2.$$

The first line is positive, while the second can produce negative depending on the residuals.

Log-Concave Coupling: Let auxiliary $\xi_{i,k}$, one for each pair of observation index i and neuron index k, be arranged to be conditionally independent with $\xi_{i,k}$ given w distributed normal with mean $x_i \cdot w_k$ and variance $1/\rho$ with $\rho = \beta CV/K$, where $C = C_{Y,V} = \max_i |Y_i| + V$ bounds the $|res_i(w)|$ for all w. Then w and ξ are coupled, having a joint density $p(w, \xi) = p(w)p(\xi|w)$, with a reverse conditional density $p(w|\xi)$ proportional to $p_0(w) \exp\{-\beta \ell_\xi(w)\}$ where $\ell_\xi(w) = \ell(w) + \frac{1}{2}\frac{V}{K} C \sum_{i=1}^n \sum_{k=1}^K (x_i \cdot w_k - \xi_{i,k})^2$. This $\ell_\xi(w)$ has Hessian $H_\xi(w)$ with a quadratic form $a^{\mathrm{T}} H_\xi(w) a$ taking the same form as above but with a new second term

$$\sum_{i=1}^n \sum_{k=1}^K \left[\tfrac{V}{K} C - c_k \, res_i(w) \psi''(x_i \cdot w_k)\right](a_k \cdot x_i)^2$$

made positive for w in S_1^K, for each ξ in R^{NK}, because $|c_k| \leq V/K$, $|res_i(w)| \leq C$, and $|\psi''| \leq 1$. So $p(w|\xi)$ is log-concave in w for each ξ, implying there are computationally rapid (low-order polynomial time in N, K, d) samplers of w given ξ, using e.g. [5].

Restrict ξ to the event that the linear combinations $\sum_{i=1} \xi_{i,k} x_{i,j}$ (for each variable j and neuron K) are in intervals centered at their mean and extending to $\sqrt{2\log(Kd)}$ standard deviations. Restricting these linear combinations to such intervals provides a convex polytope in which the auxiliary random vectors ξ reside with high probability.

Density of ξ as a Mixture and its Score: The density of ξ is $p(\xi) = \int p(w, \xi) dw$, the integral of a log-concave function. We have interest in its score $\nabla \log 1/p(\xi)$ and associated Hessian $\tilde{H}(\xi) = \nabla \nabla^{\mathrm{T}} \log 1/p(\xi)$. By the projection rule for marginal scores $\nabla \log 1/p(\xi)$ is $E[\nabla \log 1/p(\xi|w)|\xi]$, where the normal conditional score $\partial \xi_{i,k} \log 1/p(\xi|w)$ has the affine form $\rho \, \xi_{i,k} - \rho \, x_i \cdot w_k$. Accordingly $\partial \xi_{i,k} \log 1/p(\xi)$ equals $\rho \, \xi_{i,k} - \rho \, x_i \cdot E[w_k|\xi]$. Given ξ, these are efficiently computed by Monte Carlo sampling of $w|\xi$.

Stochastic Diffusion Sampling of ξ: Armed with the ξ score at each time step, we have access to (time-discretized versions) of the Langevin diffusion with gradient drift [6]

$$d\xi(\tau) = \tfrac{1}{2} \nabla \log p(\xi(\tau)) d\tau + dB(\tau).$$

Starting with Gaussian $\xi(0)$, the density of $\xi(\tau)$ converges, as time τ increases, to the invariant $p(\xi)$. The rapidity of that convergence is addressed by the log concavity of $p(\xi)$.

Log-Concavity of p(ξ): Hessian $\tilde{H}(\xi) = \nabla \nabla^{\mathrm{T}} \log 1/p(\xi) = \rho \left\{ I - \rho \, Cov \left[\begin{smallmatrix} Xw_1 \\ Xw_K \end{smallmatrix} | \, \xi \right] \right\}$ so its quadratic form for unit a in R^{nK} with blocks a_k is $a^{\mathrm{T}} \tilde{H}(\xi) a = \rho \{1 - \rho \, Var[\tilde{a} \cdot w|\xi]\}$, where $\tilde{a} = \left[\begin{smallmatrix} X' a_1 \\ X' a_K \end{smallmatrix} \right]$ with $||\tilde{a}||^2 \leq nd$, with variance using the log-concave $p(w|\xi)$. It is conjectured to be more concentrated, producing smaller variance of linear combinations, than with the prior $p_0(w)$, for which the counterpart is evaluated using $Cov_0(w_k) = \frac{2}{(d+2)(d+1)} I$. A Hölder inequality confirms that $Var[\tilde{a} \cdot w|\xi]$ is not larger than $Var_0[\tilde{a} \cdot w]$ by too large a factor.

Toward that end, let $\tilde{\ell}_\xi(w) = \ell_\xi(w) - E_0[\ell_\xi(w)]$, where we have subtracted the expectation using the prior. Restricting attention to ξ in the discussed set of high probability, it is seen that $|\tilde{\ell}_\xi(w)| \leq 9VC$ where $C = C_{Y,V}$. Let $\mu_\xi = E[\tilde{a} \cdot w|\xi]$. Then $Var[\tilde{a} \cdot w] = E_0[(\tilde{a} \cdot w - \mu_\xi)^2 \exp\{-\beta \tilde{\ell}_\xi(w) - \Gamma_\xi(\beta)\}]$ where the exponential is the likelihood factor, with log normalizer $\Gamma_\xi(\beta)$, the cumulant generating function of $-\tilde{\ell}_\xi(w)$ with respect to the prior.

Lemma: *Conditional Variance Concentration.* For any $r \geq 1$,

$$\begin{aligned} Var[\tilde{a} \cdot w|\xi] &\leq [E_0[(\tilde{a} \cdot w)^{2r}]]^{1/r} \exp\{\tfrac{r-1}{r} \Gamma_\xi(\tfrac{r}{r-1}\beta) - \Gamma_\xi(\beta)\} \\ &\leq \tfrac{4nr}{(d+2)e} \exp\{9 \, \beta \, V c \, n/r\}, \end{aligned}$$

which is $36\beta \, Vcn^2/(d+2)$ at the optimal $r = 9\beta V c n$. Thus $\rho \, Var[\tilde{a} \cdot w|\xi] \leq 36 V^2 c^2 \beta^2 n^2/(Kd)$, which is less than $1/2$ when the number of parameters Kd exceeds a multiple of $(\beta n)^2$.

Proof of Lemma: *Summary.* The variance is $E_0[(\tilde{a}\cdot w-\mu_\xi)^2 \exp\{-\beta\tilde{\ell}_\xi(w)-\Gamma_\xi(\beta)\}]$, which is increased by replacing μ_ξ with 0. The first bound is from Hölder's inequality. The second follows from evaluation of prior moments together with first-order Taylor expansion of the cummulant generating function $\Gamma_\xi(\beta)$ using the fact that it is 0 at $\beta = 0$.

Corollary: *High-Dimensional Log-Concavity of* $p(\xi)$. Thus, for ξ in the indicated high-probability set, the density $p(\xi)$ is strictly log-concave, with $\nabla\nabla \log 1/p(\xi) \geq \rho/2$ when the number of parameters Kd exceeds $72V^2c^2\beta^2n^2$.

Thus ξ can be rapidly sampled via the theory for log-concave densities as in [5],[6].

The Neural Net Posterior as a Mixture: Though $p(w)$ is not log-concave, it is $p(w) = \int p(w|\xi)p(\xi)d\xi$, a mixture of log-concave densities $p(w|\xi)$ using a $p(\xi)$ that is also log-concave, as long as the parameter dimension is sufficiently large. *Accordingly, a draw from* $p(w)$ *can be achieved via a draw from* $p(\xi)$ *followed by a draw from* $p(w|\xi)$. *This gives the polynomial-time computational feasibility of sampling from the neural net posterior* $p(\xi)$.

Statistical Risk and Arbitrary-Sequence Regret: Consider statistical risk bounds available via online learning regret. Take the (x_t, y_t) as 'time'-ordered for $t = 1, 2,...,N$. Prediction at time t is obtained by using the posterior distribution based on the preceding data $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$ using the posterior density $p(w) = p_{t-1}(w)$ given above with $n = t-1$. Its prediction for $f(x)$ at $x = x_t$ is $f_t(x) = \int f(x, w)p_{t-1}(w)dw = E_{p_{t-1}}\big[f_t(x, w)\big]$ computed by Monte Carlo sampling from $p_{t-1}(w)$ and averaging the values of $f(x, w)$. Regret definitions and their relationships are fairly standard. What's new is the application to provably computationally-feasible and accurate neural net estimates.

Individual Observation Regret: Compared to an arbitrary $g(x)$, the individual *squared error regret* at time t for $t = 1, 2, \ldots, N$ is $r_t^{square} = r_{t,g}^{square}$ given by

$$r_t^{square} = \tfrac{1}{2}[(y_t - f_t(x_t))^2 - (y_t - g(x_t))^2]$$

where $f_t(x_t) = E_{p_{t-1}}\big[f_t(x_t, w)\big]$. An obvious bound is to pull the expectation outside the square. The result is the expected individual squared error regret $r_t^{rand} = r_{t,g}^{rand}$ given by

$$r_t^{rand} = \tfrac{1}{2}[E_{p_{t-1}}[\big(y_t - f(x_t, w)\big)^2] - \big(y_t - g(x_t)\big)^2].$$

Meanwhile, let $p(y|g(x))$ be the normal density with mean $g(x)$ and variance $1/\beta$ and let $\hat{p}_t(y|x) = E_{p_{t-1}}[p(y_t|f(x_t, w))]$ be the Bayes predictive density based on the preceding data. The individual *logarithmic predictive density regret* $r_t^{log} = r_{t,g}^{square}$ is given by

$$r_t^{\log} = \tfrac{1}{\beta}\left[\log 1/\hat{p}_t(y_t|x_t) - \log 1/p(y_t|g(x_t))\right].$$

Let us suppose that, like $f(x, w)$, the $g(x)$ are bounded by a constant b. Define the error of g as $\epsilon_t = \epsilon_{t,g} = y_t - g(x_t)$ and set $c_t = c_{\epsilon_t,b} = |\epsilon_t|b + b^2$, not depending on w.

Lemma: *Comparing Individual Regret:* $r_t^{\log} \leq r_t^{rand}$ and $r_t^{square} \leq r_t^{rand} \leq r_t^{log} + 2\beta c_t^2$.

Proof of Lemma: The $r_t^{\log} \leq r_t^{rand}$ and $r_t^{square} \leq r_t^{rand}$ are by Jensen's inequality. Consider $\frac{1}{2}[(y_t-f(x_t, w))^2-(y_t-g(x_t))^2]$ with w having the density $p_{t-1}(w)$. Then r_t^{rand} is its expected value and $r_t^{\log}$ is $1/\beta$ times its cumulant generating function at β. By the difference of squares identity, it is $(\epsilon_t + (f(x_t, w) + g(x_t))/2)(f(x_t, w) - g(x_t))$ which is less that $2c_t$. By second order Taylor expansion the cumulant generating function of a bounded random variable matches its mean to within half the square of the bound. Hence $r_t^{rand} \leq r_t^{\log} + 2\beta c_t^2$.

Corresponding time-average regret quantities are $R_{N,g}^{square} = \frac{1}{N}\sum_{t=1}^N r_{t,g}^{square}$ and likewise $R_{N,g}^{rand} = \frac{1}{N}\sum_{t=1}^N r_{t,g}^{rand}$ and $R_{N,g}^{\log} = \frac{1}{N}\sum_{t=1}^N r_{t,g}^{\log}$. Also let $c_N^2 = \frac{1}{N}\sum_{t=1}^N c_{\epsilon_t,b}^2$.

Corollary: *Comparing Time-Average Regret:* $R_{N,g}^{square} \leq R_{N,g}^{rand} \leq R_{N,g}^{\log} + 2\beta\, c_N^2$.

The product of the computationally feasible predictive densities $\prod_{t=1}^N \hat{p}_t(y_t|x_t)$ represents the Bayes factor $p(y^N|x^N) = \int p(y^N|x^N, f_w)P_0(dw)$, where P_0 is the prior. It is compared with $p(y^N|x^N, g)$, the product of normals with means $g(x_t)$. By the log product rule, as in [8],[9],[10],[11], one has the following.

Lemma: *Log Bayes Factor Representation of Logarithmic Regret:* $R_{N,g}^{\log} = \frac{1}{N\beta} \log \frac{p(y^N|x^N,g)}{p(y^N|x^N)}$.

Lemma: *Empirical Resolvability Bound on Log Regret:* The $R_{N,g}^{\log}$ is not more than

$$Resolve_{N,\beta}(A, g) = \int_A \tfrac{1}{N\beta} \log \tfrac{p(y^N|x^N,g)}{p(y^N|x^N,f_w)} P_0(dw|A) + \tfrac{1}{N\beta} \log \tfrac{1}{P_0(A)}$$

for any measurable subset $A = A_g$. This empirical resolvability has the representation

$$Resolve_{N,\beta}(A, g) = \tfrac{1}{N} \int_A (\ell(f_w) - \ell(g)) P_0(dw|A) + \tfrac{1}{N\beta} \log \tfrac{1}{P_0(A)}$$

where we recall $\ell(g) = \frac{1}{2}\sum_{t=1}^N (y_i - g(x_i))^2$ is half the empirical squared error.

Proof: Reduce the Bayes factor integral by restricting it to A and then use Jensen's inequality. The squared error representation is from our choice of $p(y^N|X^n, g)$.

Implications for the Stochastic Setting: Suppose (x_t, y_t) are independent from $P_{X,Y}$ with $y|x$ having mean $f(x)$ and variance bound σ^2. Set $g = f$ and take the expected value to produce two bounds on the mean square risk of estimators of f.

Corollary: *Expected Resolvability Bound on Mean Square Risk*: For any distribution on $Y|X$ with conditional variance bounded by σ^2, any $\beta > 0$ and any choice of $A = A_f$, the mean square generalization error is bounded by

$$E\big[||\hat{\hat{f}} - f||^2\big] \le \tfrac{1}{N}\sum_{t=1}^N E\big[||\hat{f}_t - f||^2\big] \le 2\, resolve_{N\beta}(A, f) + 4\beta c_{\sigma,b}^2.$$

Here $\hat{\hat{f}}(x) = \frac{1}{N}\sum_{t=1}^N \hat{f}_t(x)$, the $c_{\sigma,b} = \sigma b + b^2$ and $resolve_{N\beta}(A, f)$ is the expected resolvability

$$resolve_{N\beta}(A, f) = \tfrac{1}{2}\int_A ||f_w - f||^2 P_0(dw|A) + \tfrac{1}{N\beta}\log\tfrac{1}{P_0(A)}.$$

If $Y|X$ is Normal$(f(X), \sigma^2)$ and $\beta = 1/\sigma^2$, using the expected log regret bound directly, yields a Kullback risk of the predictive density estimates, with no need for the $4\beta c^2$ term.

$$E\big[D(p_f||\hat{\hat{p}}_N)\big] \le \tfrac{1}{N}\sum_{t=1}^N E\big[D(p_f||\hat{p}_t)\big] \le resolve_{N/\sigma^2}(A, f).$$

Here $\hat{\hat{p}}_N(y|x) = \frac{1}{N}\sum_{t=1}^N \hat{p}_t(y|x)$ and $D(p_f||\hat{p}) = D(p_{Y|X}||\hat{p}_{Y|X})$ is the Kullback loss between the conditional density and an estimator $\hat{p}$.

Discretization: Consider a discretized uniform prior which makes $\underline{w}_k$ independent on the restriction $S_{1,M}^d$ of the simplex S_1^d to vectors with rational coordinates of denominator M. For sufficient size M and Kd, the $p(\xi)$ remains log-concave and, for the discretized log-concave $p(w|\xi)$, there are sampling strategies, such as in [7], that still mix rapidly.

Resolvability for Neural Nets: For f in the convex hull of $V\Psi$, by methods from [4][12], there are $\underline{w}_1$,...,$\underline{w}_K$ in $S_{1,M}^d$ with $||f_w - f||^2 \le \frac{V^2}{K} + \frac{V^2}{M}$. Take A_f to be a singleton at such w. There are at most $(2d)^{MK}$ points in the support of the prior. Accordingly, setting $M = K$, an expected resolvability bound is $resolve_{N\beta}(f) \le \frac{V^2}{K} + \frac{K^2}{N\beta}\log(2d)$. For this bound the optimal number of neurons is $K = V^{2/3}(N\beta)^{1/3}/(2\log 2d)^{1/3}$, at which the resolvability bound is $2.05\, V^{4/3}(\log 2d)^{1/3}/(N\beta)^{1/3}$. In the normal error model, with $\beta = 1/\sigma^2$, this provides Kullback risk of a computationally-feasible estimator of order $((\log 2d)/N)^{1/3}$.

In the general error model, with the added $4C^2\beta$, the best β is $0.5(K/C)((\log 2d)/N)^{1/2}$, and for the best K is $0.5(V/C^{1/2})(N/(\log 2d))^{1/4}$, at which the resulting computationally-feasible mean square risk bound is

$$4VC^{1/2}\left(\tfrac{\log(2d)}{N}\right)^{1/4}.$$

Neural net mean square risk with a 1/2 power (instead of 1/3 or 1/4) is available from knowledge of the metric entropy or the Gaussian complexity of the convex hull of $V\Phi$ as in [13],[14],[15]. However, these use empirical criteria with potentially computationally infeasible optimization, or Bayes posteriors as in [10] with computationally infeasible priors on optimal covers. It is not yet known if risk with the 1/2 power is computationally feasible.

References

[1] A.R. Barron, C. McDonald. "Log Concave Coupling for Sampling From Neural Net Posterior Distributions," *NUS-IMS Workshop on Statistical Machine Learning for High Dimensional Data*, stat.yale.edu/∼arb4/presentations/SingaporeLogConcaveCouplingForNeuralNets.pdf 2024.

[2] C. McDonald, A.R. Barron. "Log Concave Coupling for Sampling Neural Net Posteriors," *Proc. IEEE Internat. Symposium on Information Theory*, 2024.

[3] A.R. Barron. "Shannon Lecture: Information Theory and High-Dimensional Bayesian Computation", *IEEE Internat. Symposium on Information Theory*, stat.yale.edu/∼arb4/ShannonLecture.pdf, 2024.

[4] A.R. Barron. "Universal Approximation Bounds for Superpositions of Sigmoidal Function" *IEEE Transactions on Information Theory*, Vol.IT-39, pp.930-944, 1993.

[5] L. Lovász, S. Vempala. "The Geometry of Log Concave Functions and Sampling Algorithms," *Random Structures & Algorithms*, Vol.30, p.307-358, 2007.

[6] V. Srinivasan, A. Wibisono, A. Wilson. *Fast sampling from constrained spaces using the Metropolis-adjusted Mirror Langevin algorithm*, ArXiv:2312.08823v3, 2024.

[7] D. Applegate, R. Kannan. "Sampling and Integration of Near Log-Concave Functions," *Proc. 23rd ACM Symposium on Theory of Computing*, p.156-163, 1991.

[8] A.R. Barron. "Are Bayes Rules Consistent in Information?" *Open Problems in Communication and Computation*, Springer, p.85-91, 1987.

[9] A.R. Barron. "Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems", *Bayesian Statistics 6*, p.27-52, 1998.

[10] Y. Yang, A.R. Barron. "Information-Theoretic Determination of Minimax Rates of Convergence," *Annals of Statistics*, Vol.27,p.1564-1599, 1999.

[11] Q. Xie, A.R. Barron. "Asymptotic Minimax Regret for Data Compression, Gambling and Prediction," *IEEE Transactions on Information Theory*, Vol.46, p.431-335, 2000.

[12] S. Chatterjee, A.R. Barron. "Information Theoretic Validity of Penalized Likelihood," *Proc. IEEE Internat. Symposium Information Theory*, p.3027-3031, extended version: arXiv:1401.6714v2, 2014.

[13] J.M. Klusowski, A.R. Barron. "Approximation by Combinations of ReLU and Squared ReLU with ℓ_1 and ℓ_0 Controls", *IEEE Transactions on Information Theory*, Vol.64, p.7649-7656, 2018.

[14] A.R. Barron, J.M. Klusowski. "Approximation and Estimation for High-Dimensional Deep Learning Networks", ArXiv:1809.03090v2, 2018.

[15] A.R. Barron, J.M. Klusowski. "Complexity, Statistical Risk and Metric Entropy of Deep Nets using Total Path Variation," ArXiv:1902.00800v2, 2019.

Yale University, Department of Statistics and Data Science
Email address: Andrew.Barron@yale.edu