Approximating the Trace Norm Distribution Partition Function[*]

Jason D. M. Rennie
jrennie@gmail.com

January 30, 2006

Abstract

The Trace Norm Distribution is a distribution over matrices such that the negative log-likelihood is proportional to the trace norm of the matrix. The partition function of this distribution is intractible to compute exactly for large matrices, so we must use approximation techniques. We discuss sampling and bounding techniques that can be used to approximate the partition function.

1 Introduction

The trace norm distribution of a matrix, $X \in \mathbb{R}^{n \times m}$ (we assume $n > m$) is defined as

$$P_\lambda(X) = \frac{1}{Z_\lambda} \exp(-\lambda \|X\|_{\mathrm{tr}}), \tag{1}$$

where $\|X\|_{\mathrm{tr}}$ is the trace norm of X, the sum of (non-negative) singular values. Z_λ is the partition function of the distribution, which is the integral of $\exp(-\lambda\|X\|_{\mathrm{tr}})$ over all matrices $X \in \mathbb{R}^{n \times m}$,

$$Z_\lambda = \int_{\mathbb{R}^{n \times m}} \exp(-\lambda \|X\|_{\mathrm{tr}}) dX. \tag{2}$$

Clearly X is an awkward representation for calculating the sum of singular values of a matrix. So, we change variables to a singular-value decomposition (SVD) factorization, $X = U\Sigma V^T$. Note that doing this does not change the number of free parameters, since U has orthonormal columns, V is orthogonal and Σ is diagonal. The diagonal elements of Σ are the singular values and are ordered from largest to smallest. Edelman gives the Jacobian for the SVD change of variables [1]. Note that applying a sign change to corresponding

[*] Joint work with John Barnett and Tommi Jaakkola.

columns of U and V does not change X, so we must divide by 2^m. Our integral becomes

$$Z_\lambda = \frac{1}{2^m} \int \exp\left(-\lambda \sum_{i=1}^{m} \sigma_i\right) \prod_{i<j} (\sigma_i^2 - \sigma_j^2) \prod_{i=1}^{m} \sigma_i^{n-m} d\Sigma^\wedge (H^T dU)^\wedge (V^T dV)^\wedge, \tag{3}$$

where $H \in \mathbb{R}^{n\times n}$ is an orthogonal matrix with first m columns identical to U. This integral separates nicely into three parts, one for each part of the decomposition. $\int (H^T dU)^\wedge$ and $\int (V^T dV)^\wedge$ are integrals over the Stiefel manifold and orthogonal group, respectively. Edelman gives the formula for these volume calculations [2]. What remains is integration over the singular values,

$$\int_0^\infty \int_0^{\sigma_2} \cdots \int_0^{\sigma_{m-1}} \exp\left(-\lambda \sum_{i=1}^{m} \sigma_i\right) \prod_{i<j} (\sigma_i^2 - \sigma_j^2) \prod_{i=1}^{m} \sigma_i^{n-m} d\sigma_m \ldots d\sigma_2 d\sigma_1 \tag{4}$$

Note that $\prod_{i<j}(\sigma_i^2 - \sigma_j^2)$ expands to a polynomial with more than 2^{m-1} terms, so the integral is intractible for large m. We must turn to approximation techniques for large matrices. We describe these in the next two sections.

2 Approximation via Sampling

We would like to calculate

$$Z_\lambda = \int_{\mathbb{R}^{n\times m}} \exp(-\lambda \|X\|_{\mathrm{tr}}) dX. \tag{5}$$

But, this is intractible for large n, m. However, a technique know as importance sampling [3] allows us to estimate this quantity. Let $f(X) = \exp(-\lambda \|X\|_{\mathrm{tr}})$. Let $g(X)$ be a probability distribution over $\mathbb{R}^{n\times m}$ from which we can sample. Clearly, $Z_\lambda = \int f(X) g(X)/g(X) dX$. Let $\{X_1, X_2, \ldots, X_n\}$ be samples from g. Then,

$$Z_\lambda \approx \frac{1}{n} \sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)}. \tag{6}$$

Clearly, the approximation is best when $g \propto f$. However, if we knew such a g, sampling would not be necessary. The distribution g should be easy to compute, but should be approximately proportional to f. In our case, a reasonable choice for g is the distribution where each column of X has a negative log-probability equal to its Euclidean distance from the origin,

$$g(X) \propto \prod_{j=1}^{m} e^{-\lambda \|X_j\|}, \tag{7}$$

where X_i is the i^{th} column of X. The partition function for g is not completely trivial, so we provide its calculation.

$$\int e^{-\lambda\|\vec{x}\|} d\vec{x} = \int_0^\infty e^{-\lambda r} dr \int (H dq)^\wedge = \frac{2\pi^{n/2}}{\lambda\Gamma(\frac{n}{2})}, \tag{8}$$

where H is the householder reflector and $\vec{q}$ is the spherical parameter vector as discussed in [2]. $\int (H dq)^\wedge$ calculates the surface area of the n-sphere. Hence, $g(X) = \left(\frac{\lambda\Gamma(\frac{n}{2})}{2\pi^{n/2}}\right)^m \prod_{j=1}^m e^{-\lambda\|X_j\|}$. g is not difficult to sample from. Sample a radius r from a two-sided exponential. Sample an n-vector, $\vec{x}$, from a standard Normal. Scale $\vec{x}$ to length r to yield a sample from g.

3 Approximation via Bounding

Another approach to approximating Z_λ is to bound the integral with easy-to-compute quantities. We can easily construct an upper bound to the integral by noting the following inequalities,

$$\prod_{i<j} (\sigma_i^2 - \sigma_j^2) \le \prod_{i=1}^m \sigma_i^{2(m-i)}, \quad \text{and} \quad \int_0^x f(t) e^{-t} dt \le \int_0^\infty f(t) e^{-t} dt, \tag{9}$$

where $f(t) \ge 0$ $\forall t$ and we assume that the singular values are positive and ordered ($\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_m$). Applying these two inequalities to (4) gives us an upper bound on the integral,

$$\int_0^\infty \cdots \int_0^\infty \prod_{i=1}^m e^{-\lambda\sigma_i} \sigma_i^{n+m-2i} d\sigma_m \ldots d\sigma_1 = \prod_{i=1}^m \frac{\Gamma(n+m-2i)}{\lambda^{n+m-2i}}. \tag{10}$$

The term on the right is easy to compute since all Gamma arguments are positive integers.

Another upper bound can be derived from the sampling distribution we used above. Note that

$$e^{-\lambda\|X\|_{\text{tr}}} \le \prod_{j=1}^m e^{-\lambda\|X_j\|}, \tag{11}$$

where X_j is the j^{th} column of X. Hence,

$$Z_\lambda \le \left(\frac{\lambda\Gamma(\frac{n}{2})}{2\pi^{n/2}}\right)^m. \tag{12}$$

References

[1] A. Edelman. Jacobians of matrix transforms (with wedge products). http://web.mit.edu/18.325/www/handouts.html, February 2005. 18.325 Class Notes: Finite Random Matrix Theory, Handout #3.

[2] A. Edelman. Volumes and integration. http://web.mit.edu/18.325/www/handouts.html, March 2005. 18.325 Class Notes: Finite Random Matrix Theory, Handout #4.

[3] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.