MEAN-FIELD APPROXIMATIONS FOR LOG-CONCAVE DISTRIBUTIONS

MRIGANKA BASU ROY CHOWDHURY

1. Log-concave distributions

- A distribution P on $\mathbb{R}$ is log-concave if its density f is log-concave, i.e., $d\mu/d\lambda \propto \exp(f(x))$ where f is a concave function. These measures appear in optimization and sampling.
- We call a function κ-strongly concave if $f(x) + \kappa \left\| x \right\|^2/2$ is concave. For smooth f, this is equivalent to $\nabla^2 f \preceq -\kappa I$. The corresponding measures are called κ-strongly log-concave. For instance, the standard Gaussian is log-concave with $f(x) = -\left\| x \right\|^2/2$. In fact, it is strongly log-concave with $\kappa = 1$.
- A wide range of facts are known for such distributions. The canonical Markov process associated with these measures are called Langevin diffusions. They are reversible with respect to the measure and are defined via the SDE
$$dX_t = \nabla f(X_t)dt + \sqrt{2}dB_t,$$
- An analysis of this diffusion yields two remarkable facts (whenever the objects are defined):
  - Poincare inequality:
$$\mathbf{Var}_\mu(\phi) \leqslant \frac{1}{\kappa} \int \left\| \nabla \phi \right\|^2 d\mu. \tag{1}$$
  - Log-Sobolev inequality:
$$\mathrm{KL}\left(\nu \| \mu\right) \leqslant \frac{1}{2\kappa} \int \left\| \nabla \log \frac{d\nu}{d\mu} \right\|^2 d\nu. \tag{2}$$
- Another nontrivial fact here is that log-concave distributions are closed are marginalization.

2. Mean-field approximations

- Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is such that $f(x) = \sum_i f_i(x_i)$ for some κ-concave f_i, then so is f. The Gibbs measure $\mu \propto \exp(f)$ is then a *product law*. The question is, given a general f, how close is it to a product measure.
- This close-ness can be measured via the KL divergence. What is the product law $\nu = \otimes \nu_i$ which minimizes $\mathrm{KL}\left(\nu \| \mu\right)$?
- Let us expand this. We have
$$\begin{aligned}\mathrm{KL}\left(\nu \| \mu\right) &= \int \log \frac{d\nu}{d\mu} d\nu \\ &= \int \log \frac{d\nu}{d\mu} d\nu \\ &= \int \log d\nu - (f - \log Z) d\nu \\ &= \log Z - \left(\mathbb{E}_\nu f + H(\nu)\right).\end{aligned}$$

where $H(\nu) = -\int \log \nu \cdot d\nu$ is the differential entropy of ν. This also shows that the optimizer produces the best mean-field approximation in terms of the partition function. Recall, via a trivial reformulation, that the optimizer of

$$\mathbb{E}_\nu f + H(\nu)$$

over all ν was the Gibbs measure $\mu \propto \exp(f)$. To get a sense of $H(\nu)$, note that

$$H(N(0,\sigma^2)) = -\mathbb{E}_X\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-X^2/2\sigma^2)\right)\right] = \frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}E[-X^2] = \frac{1}{2}\log(2\pi e\sigma^2)$$

which becomes larger with increasing σ.

3. A simple heuristic

- Ignoring questions about existence and uniqueness, suppose $\nu = \otimes \nu_i$ is an optimizer of $\mathbb{E}_\nu f + H(\nu)$. Write $\nu = \nu_1 \otimes \nu_{>1}$ where $\nu_{>1} = \otimes_{i>1}\nu_i$. Then, $H(\nu) = H(\nu_1) + H(\nu_{>1})$.
- Further,
$$\mathbb{E}_\nu f = \mathbb{E}_{X\sim\nu_1, Y\sim\nu_{>1}} f(X,Y) = \mathbb{E}\left[\mathbb{E}\left[f(X,Y)|X\right]\right]$$
Thus defining
$$f_1(x) = \mathbb{E}_{Y\sim\nu_{>1}}[f(x,Y)] = \mathbb{E}_{(X,Y)\sim\nu}\mathbb{E}\left[f(X,Y)|X=x\right]$$
we have that ν_1 maximizes
$$\mathbb{E}_{\nu_1} f_1 + H(\nu_1).$$
and thus $d\nu_1/d\lambda \propto \exp(f_1)$.
- This results in the *fixed point equation* for an optimizer ν:

$$\frac{d\nu_i}{d\lambda}(x_i) \propto \exp\left(\mathbb{E}_{X\sim\nu}[f_i(X)|X_i = x_i]\right). \tag{3}$$

4. Analyzing the fixed point equation

- Let us consider a solution ν to the fixed point equation (3), and examine the consequences.
- Fix a coordinate, say 1, and consider $f_1(x_1) = \mathbb{E}_{X\sim\nu}[f(X)|X_1 = x_1]$, which appears as the log-density of ν_1. Then
$$f_1'(x_1) = \mathbb{E}_{X\sim\nu}[\partial_1 f(X)|X_1 = x_1],$$
due to the independence in ν. Another derivative yields
$$f_1''(x_1) = \mathbb{E}_{X\sim\nu}[\partial_{11} f(X)|X_1 = x_1] \leqslant -\kappa.$$
since $\partial_{11} f(X) = \left(\nabla^2 f(X)\right)_{11} \leqslant -\kappa$. Thus, f_1 is κ-strongly concave in dimension one, and thus ν_1 is κ-strongly log-concave.

- Applying the Log-Sobolev inequality in (2) to ν_1 yields

$$\begin{aligned}\mathrm{KL}\left(\nu\|\mu\right) \leqslant \frac{1}{2\kappa}\int \left\|\nabla \log \frac{d\nu}{d\mu}\right\|^2 d\nu &= \frac{1}{2\kappa}\mathbb{E}_{Y\sim\nu}\left[\sum_{i=1}^{d}\left(\mathbb{E}_{X\sim\nu}[\partial_i f(X)|X_i=Y_i]-\partial_i f(Y)\right)^2\right]\\ &= \frac{1}{2\kappa}\sum_{i=1}^{d}\mathbb{E}_{Y\sim\nu}\left[\mathbb{E}\left[\left(\mathbb{E}_{X\sim\nu}[\partial_i f(X)|X_i=Y_i]-\partial_i f(Y)\right)^2\Big|Y_i\right]\right]\\ &= \frac{1}{2\kappa}\mathbb{E}_{Y\sim\nu}\left[\sum_{i=1}^{d}\mathbf{Var}_{X\sim\nu}[\partial_i f(X)|X_i=Y_i]\right]\end{aligned}$$

At this point, observe that since each factor of ν_i is κ-strongly log-concave, so is $\otimes_{k\neq i}\nu_k$. Therefore, the Poincare inequality (1) is in play, yielding

$$\mathbf{Var}_{X\sim\nu}[\partial_i f(X)|X_i=Y_i] \leqslant \frac{1}{\kappa}\mathbb{E}_{X\sim\nu}\left[\sum_{j\neq i}(\partial_{ij}f(X))^2\Big|X_i=Y_i\right].$$

Putting things together, we get

$$\begin{aligned}\mathrm{KL}\left(\nu\|\mu\right) &\leqslant \frac{1}{2\kappa^2}\mathbb{E}_{Y\sim\nu}\left[\sum_{i=1}^{d}\mathbb{E}_{X\sim\nu}\left[\sum_{j\neq i}(\partial_{ij}f(X))^2\Big|X_i=Y_i\right]\right]\\ &= \frac{1}{\kappa^2}\sum_{i<j}\mathbb{E}_{Y\sim\nu}\left[\mathbb{E}_{X\sim\nu}[(\partial_{ij}f(X))^2|X_i=Y_i]\right]\\ &= \frac{1}{\kappa^2}\sum_{i<j}\mathbb{E}_{X\sim\nu}[(\partial_{ij}f(X))^2].\end{aligned}$$

- Recalling that this KL divergence is also the gap in the Gibbs variational principle, this shows that there is a product law ν satisfying

$$0 \leqslant \log Z - (\mathbb{E}_\nu f + H(\nu)) = \mathrm{KL}\left(\nu\|\mu\right) \leqslant \frac{1}{\kappa^2}\sum_{i<j}\mathbb{E}_{X\sim\nu}[(\partial_{ij}f(X))^2].$$

- As proved in the main paper, this optimizer exists, invoking the weak-closedness of the set of product measures. We will not pursue these details here.

5. Applications

- Firstly observe that if f is linearly separable, i.e., μ is already product, $\partial_{ij}f=0$ for all $i\neq j$. Hence, the error is zero here, as should be the case.
- As a broad application, let us consider a general class of f as follows:

$$f(x)=\sum V(x_i)+\sum_{i<j}J_{ij}K(x_i-x_j),$$

for a κ-concave potential V, fixed non-negative, doubly stochastic interaction matrix J and an even concave kernel K. The potential V will serve to "confine" the measure, and the K and J produce an interaction between the coordinates.

- Firstly, this is κ-concave. To see this, observe that the Hessian of the potential part itself is $\preceq -\kappa I$, so it suffices to show that everything else is negative-definite, for which, in-turn, it suffices to show that $(x, y) \mapsto K(x-y)$ is concave if K is. For this, note that
$$\nabla^2 K(x-y) = K''(x-y) \cdot \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$
The last matrix has eigenvalues 0 and 2, so the Hessian is negative-definite, since $K'' \leqslant 0$.
- Then, for $a < b$,
$$\begin{aligned} \partial_{ab} f(x) &= \sum_{i<j} J_{ij} \partial_{ab} K(x_i - x_j) \\ &= -J_{ab} K''(x_a - x_b) \end{aligned}$$
since
$$\begin{aligned} \partial_{ab} K(x_i - x_j) &= \partial_a \left(\partial_b K(x_i - x_j)\right) \\ &= \partial_a \left(\delta_{ib} K'(x_i - x_j) - \delta_{jb} K'(x_i - x_j)\right) \\ &= \delta_{ia}\delta_{ib} K''(x_i - x_j) - \delta_a \delta_{ib} K''(x_i - x_j) - \delta_{ia}\delta_{jb} K''(x_i - x_j) + \delta_{ja}\delta_{jb} K''(x_i - x_j) \\ &= K''(x_i - x_j)\left(\delta_{ia}\delta_{ib} - \delta_{ja}\delta_{ib} - \delta_{ia}\delta_{jb} + \delta_{ja}\delta_{jb}\right), \end{aligned}$$
which is zero when $a < b$ except when $a = i, b = j$ in which case its $-K''(x_a - x_b)$.
- Thus, the gap is immediately seen to be bounded by
$$\frac{1}{\kappa^2} \sum_{a<b} J_{ab}^2 \cdot \left\|K''\right\|_\infty^2 \leqslant \frac{1}{2\kappa^2} \cdot \operatorname{tr}(J^2) \cdot \left\|K''\right\|_\infty^2 .$$
If $\kappa = \Theta(1)$, K'' is bounded and $\operatorname{tr}(J^2) = o(n)$, the error is $o(n)$. But why is n the right scale? To see this, let us compute the "mean-field partition function", i.e., the supremum of the following quantity over all product $\nu = \otimes \nu_i$:
$$\mathbb{E}_\nu f + H(\nu) = \sum_{i=1}^n \mathbb{E}_{\nu_i} V + \sum_{i<j} J_{ij} \cdot \mathbb{E}_{(X,Y)\sim\nu_i\otimes\nu_j} K(X-Y) + \sum_{i=1}^n H(\nu_i).$$
It turns out that one can prove the following lemma:

Lemma 1. *Let $\mu_1, \ldots, \mu_n$ be n distributions on $\mathbb{R}$. Then, there is a random vector X with marginals ν_i, such that*
$$H(\overline{X}) \geqslant n^{-1} \sum_{i=1}^n H(\mu_i)$$
where $\overline{X}$ is the average of the entries of X.

This, and the concavity of V, K allows one to reduce the optimization problem above to a one-dimensional problem, yielding that the optimizer is i.i.d., with marginals $\nu_i = \nu^*$, say,
$$\mathbb{E}_\nu f + H(\nu) = n\left(\int V d\nu^* + \frac{1}{2}\int K(x-y) d\nu^*(x) d\nu^*(y) + H(\nu^*)\right)$$
The inner problem has no dependence on n, and thus the quantity is $\Theta(1)$, making the entire quantity $\Theta(n)$. This shows the requirement of $o(n)$ for the error.

- As a concrete instance, consider a Hamiltonian similar to what was considered last time, only this time we will make things "spherical". Choose $V(x) = -x^2/2$, so that $\kappa = 1$, and choose $K(x) = -x^2$. Further suppose, $J = A/d$ where A is the adjacency matrix of a d-regular graph. Then,
- Then, the overall Hamiltonian looks like
$$f(x) = -\left(\frac{1}{2}\sum x_i^2 + \frac{1}{2d}\sum_{(i,j)\in E}(x_i - x_j)^2\right).$$
This is essentially n Gaussians on the vertices, weighted by the "interaction" which favors similar Gaussians on nearby vertices. This model is basically the GFF on this graph.
- The mean-field approximation is active when $\mathrm{tr}(J^2) = o(n)$. In terms of the adjacency matrix, the condition is therefore $\mathrm{tr}(A^2) = o(nd^2)$. Since $\mathrm{tr}(A^2) = |E| = O(nd)$, this holds if $d \to \infty$, the usual "mean-field" condition.
- Interestingly, in this case, the actual geometry of the regular graph does not stop the mean-field solution being i.i.d.

6. Appendix

- We deliberate on the proof of the lemma. The general fact that is true, is that for a collection of reals $t_1, \ldots, t_n$ and measures μ_n, the analogous result holds with $1/n$ replaced by t_i.
- Let us look at $n = 2$, so that we have two measures on $\mathbb{R}$, μ, ν, and we want to exhibit a coupling. The point is that $-H$ is *displacement convex* in 2-Wasserstein space. That means, that along a Wasserstein geodesic from μ to ν, say μ_t, the function $t \mapsto -H(\mu_t)$ is convex.

  Further, from Brenier, we know that the geodesic is simply $((1-t)\mathrm{id} + tT)_{\#}\mu$ for an optimal map $T : \mathbb{R} \to \mathbb{R}$ such that $T_{\#}\mu = \nu$. Then consider the coupling $(X, T(X))$ produced by Brenier. Displacement convexity finishes the proof.