Electronic Journal of Statistics
Vol. 9 (2015) 1727–1732
ISSN: 1935-7524
DOI: 10.1214/15-EJS1015

Discussion of "Hypothesis testing by convex optimization"*

Alekh Agarwal

Microsoft Research New York
641 Avenue of the Americas
New York NY 10011
e-mail: alekha@microsoft.com

MSC 2010 subject classifications: 62F03.
Keywords and phrases: Hypothesis testing, convex optimization.

Received February 2015.

It is my pleasure to congratulate the authors for this insightful and innovative piece of work. Goldenshluger, Juditsky and Nemirovski (henceforth GJN) have proposed a unifying view on a broad family of hypothesis testing problems. They consider a composite hypothesis testing problem where the goal is to identify which of two convex sets do the parameters of the distribution lie in. Remarkably, the authors provide a set of conditions under which this composite testing problem boils down to a simple test between two appropriately chosen parameters, one from each set. The authors establish near-optimality guarantees for their procedure under favorable conditions. Furthermore, the underlying computation for obtaining the test can be cast as a convex optimization problem of a form for which efficient solvers are often available. This leads to a rather comprehensive solution, in both statistical and computational terms, of a class of hypothesis testing problems.

1. Setup

The authors consider the usual setting of parametric hypothesis testing, where we are given a family of probability distributions $P_\mu \in \mathcal{P}$ over a sample space Ω, indexed by a parameter $\mu \in \mathcal{M}$. Under the assumptions, P_μ is further constrained to be an exponential family distribution, with μ as the natural parameter. Recall that this means that the density of the distribution P_μ, denoted by p_μ can be written in the form $p_\mu(\omega) = \exp(\langle \mu, \omega \rangle - A(\mu))$. Here $A(\mu)$ is the log-partition function of the exponential family, and is convex in the parameter μ [1]. Finally, the authors assume that the cumulant generating function of the distribution P_μ has a favorable structure. Specifically, they assume that for every linear operator ϕ on the sample space Ω, the function $F_\phi(\mu) = \ln(\int_\Omega \exp(\phi(\omega)) p_\mu(\omega) P(d\omega))$ is a concave function of the parameter μ.

*Main article 10.1214/15-EJS1054.

To gain some intuition, if p_μ is the Normal distribution with mean μ and identity covariance, then $F_\phi(\mu)$ is easily seen to be linear (and hence concave) in μ. When all the assumptions are satisfied, the authors refer to the setup as a *good observation scheme*.

While the last assumption might seem restrictive in general beyond the Normal distribution, the authors further establish it for the case of discrete, multinomial distributions when the parameter is the probability vector as well as for the Poisson distribution with an unknown rate parameter. They also demonstrate that a product of good observation schemes is good as well. This is important to extend their setup to multiple independent observations. Using these as building blocks, the authors further proceed to various multiple hypothesis testing settings. However, we will focus our attention on the results and consequences in pairwise hypothesis testing only for this discussion.

2. Saddle-point formulation and statistical guarantees

The authors consider a hypothesis test between the hypotheses $H_X : \mu \in X$ and $H_Y : \mu \in Y$, where $X, Y \subseteq \mathcal{M}$. They define the function

$$\Phi(\phi, [x; y]) = \ln\left(\int_\Omega \exp(-\phi(\omega))p_x(\omega)P(d\omega)\right) + \ln\left(\int_\Omega \exp(\phi(\omega))p_y(\omega)P(d\omega)\right). \tag{1}$$

Then the prescribed test is given by the solution ϕ_* to the saddle-point problem

$$2\ln(\epsilon_\star) = \min_{\phi\in\mathcal{F}} \max_{(x,y)\in X\times Y} \Phi(\phi, [x; y]).$$

Under the assumptions, it is easily seen that the objective Φ is convex in the first argument ϕ and concave in the second argument (x, y). Consequently the saddle-point exists under mild regularity conditions, and efficient algorithms to numerically compute it exist. Furthermore, $2\epsilon_\star$ yields an upper bound on the risk (which is defined to be the sum of Type-I and Type-II errors) of the test that picks H_X when $\phi_\star(\omega) \geq 0$ and H_Y otherwise.

The authors further argue that there can never be another test with a risk substantially better than that $\epsilon_\star$ (specifically, $\epsilon_\star \leq 2\sqrt{\epsilon(1-\epsilon)}$ if another test has risk of 2ϵ). Note that this seems like a potentially substantial loss in efficiency if an extremely small risk is desired since we cannot rule out the existence of another test with a risk of $\epsilon_\star^2/2$. However, the more interesting question is how many samples are required in order to reduce the risk to a pre-defined level of, say $\epsilon_0 = 0.05$. Since the risk of the test using n samples falls as $\epsilon_\star^n$, we observe that $n = \frac{\log(1/\epsilon_0)}{\log(1/\epsilon_\star)}$ samples suffice. Consequently, any other test can improve by only at most a constant factor in the number of samples required to attain a risk of ϵ_0.

Alternatively, one can also consider an indexed family of sets X_n, Y_n which are increasingly harder to distinguish, and study the smallest *separation* at which the procedure still yields a test with a risk of at most ϵ_0. Since $\epsilon_\star$ itself

is a function of n now, one obtains $n \log \frac{1}{\epsilon_\star(n)} = \log \frac{1}{\epsilon_0}$. For instance, using the expression for $\epsilon_\star$ in the case of Gaussian observation model, one sees that this yields $n\delta_n^2 = 8 \log \frac{1}{\epsilon_0}$, where $\delta_n^2 = \mathrm{dist}^2(X_n, Y_n)$ is the squared distance between the two convex sets of parameters. This is of course in line with expectation since the test is indeed optimal in the Gaussian observation model.

3. Extensions to sub-Gaussian distributions

Clearly, the GJN framework yields a very comprehensive theory in the case where all the assumptions are satisfied. It is worth asking if there are ways to extend the machinery and obtain potentially weaker results when the assumptions are not fully met. With that viewpoint, we now consider the case of exponential families, which might not satisfy the concavity assumption on $F_\phi(\mu)$. We begin by closely inspecting the form of $\Phi(\phi, [x; y])$ for general exponential families. Let $\phi(\omega) = \langle \phi, \omega \rangle + a$. It is easily seen that

$$\begin{aligned}
\ln \left(\int_\Omega \exp(\phi(\omega)) p_\mu(\omega) P(d\omega) \right) &= \ln \left(\int_\Omega \exp(\langle \phi, \omega \rangle + a + \langle \mu, \omega \rangle - A(\mu)) P(d\omega) \right) \\
&= \ln \left(\int_\Omega \exp(\langle \phi + \mu, \omega \rangle - A(\mu)) P(d\omega) \right) + a \\
&= \ln \left(\exp(A(\phi + \mu) - A(\mu)) \right) + a \\
&= A(\phi + \mu) - A(\mu) + a.
\end{aligned}$$

This immediately enables us to rewrite

$$\Phi(\phi, [x; y]) = A(-\phi + x) - A(x) + A(\phi + y) - A(y). \tag{2}$$

Crucially, the value of Φ does not depend on a like noted by GJN, which is a flexibility we will need in the sequel. Since A is a convex function of its argument, we now see that Φ is always convex in ϕ, but not necessarily concave in x, y. However, it is possible to derive upper and lower bounds on Φ with the desired properties, when the function A satisfies additional structural conditions. For the remainder, let us assume that A has L-Lipschitz continuous gradients meaning there is a norm $\|\cdot\|$ such that

$$\|\nabla A(\mu) - \nabla A(\mu')\|_* \leq L \|\mu - \mu'\| \quad \forall \mu, \mu' \in \mathcal{M}, \tag{3}$$

where $\|\cdot\|_*$ is the dual norm to $\|\cdot\|$. For some of the arguments, we will also assume that A is also λ-strongly convex with respect to the norm $\|\cdot\|$, meaning that

$$A(\mu') \geq A(\mu) + \langle \nabla A(\mu), \mu' - \mu \rangle + \frac{\lambda}{2} \|\mu - \mu'\|^2 \quad \forall \mu, \mu' \in \mathcal{M}. \tag{4}$$

When concerning ourselves with testing a particular pair H_X versus H_Y, we will need these assumptions to hold only over $X \cup Y \subseteq \mathcal{M}$, which can be important if the subsets are bounded but the entire parameter space is not. For bounded parameter sets, the smoothness condition (3) is nearly equivalent to A being

differentiable, while strong convexity of the log-partition function is typically required to avoid degeneracies in estimation and testing.

Given these conditions, we can define two auxilliary functions:

$$\overline{\Phi}(\phi, [x; y]) = \langle \nabla A(y) - \nabla A(x), \phi \rangle + L \|\phi\|^2 \quad \text{and,}$$
$$\underline{\Phi}(\phi, [x; y]) = \langle \nabla A(y) - \nabla A(x), \phi \rangle + \lambda \|\phi\|^2 . \tag{5}$$

The function $\overline{\Phi}$ is obtained from Φ by applying Equation 3 to the definition (2) twice, once with $\mu' = x - \phi, \mu = x$ and once with $\mu' = \phi + y, \mu = y$. Consequently, $\overline{\Phi}$ provides an upper bound on Φ for all tuples $(\phi, [x; y])$. Similarly, it is seen that $\underline{\Phi}$ provides a lower bound on Φ via an invocation of the strong convexity property (4).

Furthermore, both $\overline{\Phi}$ and $\underline{\Phi}$ are linear in the mean parameters $\nabla A(x)$ and $\nabla A(y)$, while the overall function is still convex in ϕ. Let us further assume that the convex sets X and Y also induce convex sets of mean parameters under the gradient mapping. Indeed, ∇A being a maximal monotone operator maps convex sets into convex sets under suitable regularity conditions (see e.g. [2]). Alternatively, one can directly consider hypotheses defined on convex sets of mean parameters, which of course uniquely identify the parameter x under the Lipschitz continuity assumption (3). Either way, we now obtain functions $\overline{\Phi}$ and $\underline{\Phi}$ which satisfy the conditions posited by GJN, and sandwich the actual Φ function. We now discuss some properties of these functions.

Proposition 1. *The functions $\overline{\Phi}, \underline{\Phi}$ defined in Equation 5 satisfy the following properties:*

1. *Both the functions $\overline{\Phi}$ and $\underline{\Phi}$ posses saddle-points, so long as the domains X, Y are convex and compact. We will denote the saddle point values as $2\ln(\overline{\epsilon})$ and $2\ln(\underline{\epsilon})$ respectively.*
2. *Let $(\phi_*, [x_*; y_*])$ be the saddle-point of $\overline{\Phi}$. Then the risk of the detector ϕ_* in testing the composite hypotheses H_X and H_Y is bounded by $2\overline{\epsilon}$.*
3. *Suppose there is another test which has a risk at most 2ϵ for testing H_X versus H_Y. Then $\underline{\epsilon} \leq 2\sqrt{\epsilon(1-\epsilon)}$.*

The proof of the proposition largely follows from easy modifications of the arguments in the proof of Theorem 2.1 in the paper.

Proof.

1. The first claim is immediate from definitions and the arguments for the saddle-points of Φ under the assumptions of GJN.
2. The proof largely follows that of Theorem 2.1.(i), along with the fact that $\overline{\Phi}$ is an upper bound on Φ. We first note that Φ, as well as $\overline{\Phi}$ are both invariant to the linear term a in the detector ϕ_*. Consequently, we can assume that it a is such that

$$\langle \nabla A(x), -\phi \rangle - a = \langle \nabla A(y), \phi \rangle + a.$$

This also means that we have

$$\ln(\overline{\epsilon}) = \langle \nabla A(x), -\phi \rangle - a + \frac{L}{2} \|\phi\|^2 = \langle \nabla A(y), \phi \rangle + a + \frac{L}{2} \|\phi\|^2 .$$

Based on this observation, it is easy to establish an analog of Lemma A.1 in our setting. Indeed, for any parameter $x \in X$, we have

$$\begin{aligned} 2\ln(\overline{\epsilon}) &= \overline{\Phi}(\phi_*, [x_*; y_*]) \geq \overline{\Phi}(\phi_*, [x; y_*]) \\ &= \langle -\nabla A(x) + \nabla A(y_*), \phi_* \rangle + L \|\phi_*\|^2 \\ &= \langle -\nabla A(x), \phi_* \rangle - a + \frac{L}{2} \|\phi_*\|^2 + \ln(\overline{\epsilon}) \\ &\geq A(x - \phi_*) - A(x) - a + \ln(\overline{\epsilon}) \\ &= \ln \left(\int_\Omega \exp(-\phi_*(\omega)) p_x(\omega) P(d\omega) \right) + \ln(\overline{\epsilon}). \end{aligned}$$

This immediately yields that the risk of ϕ_* of detecting H_Y when the parameter is drawn from the set X is at most $\overline{\epsilon}$. The claim for misdetection of H_X has an analogous proof.

3. For the third claim, we can prove an analog of Lemma A.2 from the paper. Let $(\underline{\phi}, [\underline{x}; \underline{y}])$ be a saddle-point of $\underline{\Phi}$. Note that if we have a test for distinguishing H_X from H_Y with a risk at most 2ϵ, then it is also a test for the simple hypotheses $A : \omega \sim p_{\underline{x}}$ versus $B : \omega \sim p_{\underline{y}}$. Let $p = p_{\underline{x}}$ and $q = p_{\underline{y}}$. Then the existence of such a test implies that the affinity between p and q is at most 2ϵ. On the other hand, we have

$$2\ln(\underline{\epsilon}) = \underline{\Phi}(\underline{\phi}, [\underline{x}; \underline{y}]) = \min_\phi \underline{\Phi}(\phi, [\underline{x}; \underline{y}]) \leq \min_\phi \Phi(\phi, [\underline{x}; \underline{y}]).$$

As before, the best test under Φ for testing $\underline{x}$ versus $\underline{y}$ is the likelihood ratio test, so that

$$2\ln(\underline{\epsilon}) \leq \min_\phi \Phi(\phi, [\underline{x}; \underline{y}]) = 2\ln \left(\int_\Omega \sqrt{p(\omega) q(\omega)} P(d\omega) \right).$$

Simplifying as before now yields the desired conclusion. □

Overall, the proposition shows that it is still possible to use the saddle-point machinery of GJN in order to obtain a test with an upper bound on its risk, and a lower bound on the risk of the best possible test for the composite hypotheses. Specifically, the above discussion yields a detector ϕ_* with a risk at most $2\overline{\epsilon}$, and no detector with a risk smaller than $\underline{\epsilon}^2/2$ exists. If the values $\overline{\epsilon}$ and $\underline{\epsilon}$ are not too far from each other, this is not too far from the situation guaranteed by the results of GJN.

It is worth examining some more properties of the detector ϕ_* that is obtained as a saddle-point of $\overline{\Phi}$. Note that ϕ_* is no longer a simple test between p_{x_*} and p_{y_*}. If the norm $\|\cdot\|$ used in upper and lower bounding Φ is a Euclidian norm, then ϕ_* corresponds to a simple test between Gaussians with means $\nabla A(x_*)$

and $\nabla A(y_*)$, with the covariance of the Gaussian depending on the norm. It would be tempting to consider using the optimal detector for testing p_{x_*} and p_{y_*} instead of ϕ_*. The risk of that test might be significantly larger on other parameters however. It would also be natural to successively refine the upper and lower bounds and obtain a sequence of better tests and tighter estimates of their risks. It is, however, not immediate how the argument above can be iterated in natural ways.

In summary, the work of GJN provides a general framework, within which one can study several common hypothesis testing problems. This work should open avenues for future research in relaxing the assumptions and extending similar ideas to broader hypothesis testing problems. A common situation which would be important to study in future research would be to consider nested alternatives. Concretely, it seems natural to consider tests for $H_X : x \in X$ versus $H_0 : \text{dist}(x, X) \geq \epsilon$. This situation is not easily handled in the current framework as the hypothesis H_0 is not testing for membership in a convex subset of the parameter space. Extending the framework to these cases, particularly when the set X corresponds to a subset of variables being zero would enable variable selection, for instance. Overall, we should be thankful to Goldenshluger, Juditsky and Nemirovski for their innovative work and the exciting line of questions and possibilities that it has raised for future research.

References

[1] Lauritzen, S. L., *Graphical Models*. Oxford University Press, Oxford, 1996. MR1419991

[2] Rockafellar, R. T., On the virtual convexity of the domain and range of a nonlinear maximal monotone operator. *Mathematische Annalen*, 185(2):81–90, 1970. MR0259697