JMLR: Workshop and Conference Proceedings vol 49:1–28, 2016

Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning

Ahmed El Alaoui* ELALAOUI@BERKELEY.EDU
Xiang Cheng* X.CHENG@BERKELEY.EDU
Aaditya Ramdas*,† ARAMDAS@BERKELEY.EDU
Martin J. Wainwright*,† WAINWRIG@BERKELEY.EDU
Michael I. Jordan*,† JORDAN@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences**, Department of Statistics*†*, UC Berkeley.*

Abstract

Given a weighted graph with N vertices, consider a real-valued regression problem in a semi-supervised setting, where one observes n labeled vertices, and the task is to label the remaining ones. We present a theoretical study of ℓ_p-based Laplacian regularization under a d-dimensional geometric random graph model. We provide a variational characterization of the performance of this regularized learner as N grows to infinity while n stays constant; the associated optimality conditions lead to a partial differential equation that must be satisfied by the associated function estimate $\widehat{f}$. From this formulation we derive several predictions on the limiting behavior the function $\widehat{f}$, including (a) a phase transition in its smoothness at the threshold $p = d + 1$; and (b) a tradeoff between smoothness and sensitivity to the underlying unlabeled data distribution P. Thus, over the range $p \leq d$, the function estimate $\widehat{f}$ is degenerate and "spiky," whereas for $p \geq d + 1$, the function estimate $\widehat{f}$ is smooth. We show that the effect of the underlying density vanishes monotonically with p, such that in the limit $p = \infty$, corresponding to the so-called Absolutely Minimal Lipschitz Extension, the estimate $\widehat{f}$ is independent of the distribution P. Under the assumption of semi-supervised smoothness, ignoring P can lead to poor statistical performance; in particular, we construct a specific example for $d = 1$ to demonstrate that $p = 2$ has lower risk than $p = \infty$ due to the former penalty adapting to P and the latter ignoring it. We also provide simulations that verify the accuracy of our predictions for finite sample sizes. Together, these properties show that $p = d + 1$ is an optimal choice, yielding a function estimate $\widehat{f}$ that is both smooth and non-degenerate, while remaining maximally sensitive to P.

Keywords: ℓ_p-based Laplacian regularization; semi-supervised learning; asymptotic behavior; geometric random graph model; absolutely minimal Lipschitz extension; phase transition.

1. Introduction

Semi-supervised learning is a research field of growing interest in machine learning. It is attractive due to the availability of large amounts of unlabeled data, and the growing desire to exploit it in order to improve the quality of predictions and inference in downstream applications. Although many proposed methods have been successful empirically, a formal understanding of the pros and cons of different semi-supervised methods is still incomplete.

The goal of this paper is to study the tradeoffs between some recently proposed Laplacian regularization algorithms for graph-based semi-supervised learning. In the noiseless setting, the problem amounts to a particular form of interpolation of a graph-based function. More precisely, consider a graph $G = (V, E, w)$ where $V = \{v_1, \cdots, v_N\}$ is a set of N vertices, and E is the set of edges equipped with a set $w = (w_e)_{e \in E}$ of non-negative edge weights. For some subset $O \subset V$ of the vertex set, say with cardinality $|O| = n \ll N$, and an unknown function $f^* : V \to \mathbb{R}$, suppose that we are given observations $(y_i = f^*(v_i))_{i \in O}$ of the function at the specified subset of vertices. Our goal is to use the observed values to make predictions of the function values at the remaining vertices in a way that agrees with f^* as much as possible.

In order to render this problem well-posed, the behavior of the function f^* must be tied to the properties of the graph G. In a statistical context, one such requirement is such that the marginal distribution of the points v_i be related to the regression function f^*—for instance, by requiring that f^* be smooth on regions of high density. This assumption and variants thereof are collectively referred to as the *cluster assumption* in the semi-supervised learning literature.

Under such a graph-based smoothness assumption, one reasonable method for extrapolation is to penalize the change of the function value between neighboring vertices while agreeing with the observations. A widely used approach involves using the ℓ_2-based Laplacian as a regularizer; doing so leads to the objective

$$\min_f \sum_{ij \in E} w_{ij} \big(f(v_i) - f(v_j)\big)^2 \quad \text{subject to } f(v_i) = y_i \text{ for all } i \in O, \tag{1}$$

where the penalization is enforced by the quadratic form given by the graph Laplacian (Zhu et al., 2003). This method is closely tied to heat diffusion on the graph and has a probabilistic interpretation in terms of a random walk on the graph. Unfortunately, solutions of this objective are badly behaved in the sense that they tend to be constant everywhere except for the points $\{v_i\}_{i \in O}$ associated with observations (Nadler et al., 2009). The solution must then have sharp variations near those points in order to respect the measurement constraints.

Given this undesirable property, several alternative methods have been proposed in recent work (e.g., Alamgir and Luxburg (2011); Bridle and Zhu (2013); Zhou and Belkin (2011); Kyng et al. (2015)). One such formulation is based on interpolating the observed points exactly while penalizing the maximal gradient value on neighboring vertices:

$$\min_f \max_{ij \in E} w_{ij} \left| f(v_i) - f(v_j) \right| \quad \text{subject to } f(v_i) = y_i \text{ for all } i \in O. \tag{2}$$

Any solution to this variational problem is known as an *inf-minimizer*. Kyng et al. (2015) recently proposed a fast algorithm for solving the optimization problem (2). In fact, their algorithm finds a specific solution that not only minimizes the maximum gradient value, but also the second largest one among all minimizers of the former and so on—that is to say, they find a solution such that the gradient vector $(w_{ij} \left| f(v_i) - f(v_j) \right|)_{(i,j) \in E}$ is minimal in the lexicographic ordering, which they refer to as the *lex-minimizer*. They observed empirically that when $|V| = N$ grows to infinity while both the degree of the graph and number of observations are held fixed, the lex-minimizer is a better behaved solution than its 2-Laplacian counterpart. More precisely, the observed advantage is twofold: (a) the solution is a better interpolation of the observed values; and (b) the average ℓ_1-error $\frac{1}{n} \sum_{i=1}^n |f(v_i) - f^*(v_i)|$ for the lex-minimizer remains stable, while it quickly diverges with N when f is the 2-Laplacian minimizer. Their experiments together with the known limitations of the 2-Laplacian regularization method point to the possible superiority of the ℓ_∞-based formulation (2) over the ℓ_2-based (1) in a semi-supervised setting. However, we currently lack a theoretical understanding of this assertion. Accordingly, we aim to fill this gap in the theoretical understanding of

Laplacian-based regularization by studying both formulations in the asymptotic limit as the graph size goes to infinity.

We conduct our investigation in the context of a more general objective that encompasses the approaches (1) and (2) as special cases. In particular, for a positive integer $p \geq 2$, we consider the variational problem

$$J_p(f) = \sum_{ij \in E} w_{ij}^p \, |f(v_i) - f(v_j)|^p \, . \tag{3}$$

The objective J_p is referred to as the (discrete) p-Laplacian of the graph G in the literature; for instance, see the papers by Zhou and Schölkopf (2005) and Bühler and Hein (2009), as well as references therein. It offers a way to interpolate between the 2-Laplacian regularization method and the inf-minimization approach. It is then natural to consider the general family of interpolation problems based on p-Laplacian regularization—namely

$$\min_f \; J_p(f) \quad \text{subject to } f(v_i) = y_i \text{ for } i \in O. \tag{4}$$

Formulations (1) and (2) are recovered respectively with $p = 2$ and $p \to \infty$. Indeed, in the latter case, observe that $\lim_{p\to\infty} J_p(f)^{1/p} = \max_{ij \in E} w_{ij} \, |f(v_i) - f(v_j)|$. Moreover, under certain regularity assumptions (Egger and Huotari (1990); Kyng et al. (2015)), it follows that the lex-minimizer is the limit of the (unique) minimizers of J_p as p grows to infinity[1]—that is, we have the equivalence

$$f_{\text{lex}} = \lim_{p\to\infty} \; \arg\min_u \; J_p(u) \qquad \text{subject to } u(v_i) = y_i \text{ for all } i \in O. \tag{5}$$

Our contributions: We analyze the behavior of p-Laplacian interpolation when the underlying graph G is drawn from a geometric random model. Our first main result is to derive a variational problem that is the almost-sure limit of the formulation (4) in the asymptotic regime when the size of the graph grows to infinity. For a twice differentiable function f, we use ∇f and $\nabla^2 f$ to denote its gradient and Hessian, respectively. Letting μ denote the density of the vertices in a latent space, we show that for any even integer $p \geq 2$, solutions of this variational problem must satisfy the partial differential equation

$$\Delta_2 f(x) + 2\langle \nabla \log \mu(x), \nabla f(x) \rangle + (p-2)\Delta_\infty f(x) = 0,$$

where $\Delta_2 f := \mathsf{Tr}(\nabla^2 f)$ is the usual 2-*Laplacian operator*, while $\Delta_\infty f := \frac{\langle \nabla f, \, \nabla^2 f \; \nabla f \rangle}{\langle \nabla f, \nabla f \rangle}$ is the ∞-*Laplacian operator,* which is defined to be zero when $\nabla f = 0$.

This theory then yields several predictions on the behavior of these regularization methods when the number of labeled examples is fixed while the number of unlabeled examples becomes infinite: first, the method leads to degenerate solutions when $p \leq d$; i.e., they are discontinuous, a manifestation of the curse of dimensionality. On the other hand, the solution is continuous when $p \geq d + 1$. Second, the solution *is* dependent on the underlying distribution of the data for all finite values of p; however, when $p = \infty$, the solution *is not* dependent on the underlying density μ. Consequently, as the graph size increases, the lex- and inf-minimizers end up interpolating the observed values

[1] The construction of this sequence of minimizers is known as *the Pólya algorithm*, and the study of its rate of convergence is a classical problem in approximation theory (Darst et al. (1983); Egger and Taylor (1987); Legg and Townsend (1989); Egger and Huotari (1990)).

without exploiting the additional knowledge of the density μ of the features that is provided by the abundance of unlabeled data.

In order to illustrate the consequences of this last property, we study a simple one-dimensional regression problem whose intrinsic difficulty is controlled by a parameter $\epsilon > 0$. We show that the 2-Laplacian method has an estimation rate independent of ϵ while the infinity-minimization approach has a rate that is increasing in $1/\epsilon$. As shown by our analysis, this important difference can be traced back to whether or not the method leverages the knowledge of μ. We also provide an array of experiments that illustrate some pros and cons of each method, and show that our theory predicts these behaviors accurately. Overall, our theory lends support to using intermediate value of p that will lead to non-degenerate solutions while remaining sensitive to the underlying data distribution.

2. Generative Model

We follow the popular assumption in the semi-supervised learning literature that the graph represents the metric properties of a cloud point in d-dimensional Euclidean space (Zhu et al. (2003); Bousquet et al. (2003); Belkin and Niyogi (2004); Hein (2006a); Nadler et al. (2009); Zhou and Belkin (2011)). More precisely, suppose that we are given a probability distribution on the unit hypercube $[0,1]^d$ having a smooth density μ with respect to the Lebesgue measure, as well as a bounded decreasing function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\lim_{z\to\infty} \varphi(z) = 0$. We then draw an i.i.d. sequence $(x_i)_{i=1}^N$ of samples from μ; these vectors will be identified with the vertices of the graph G: $x_i \equiv v_i$. Finally, we associate to each pair of vertices the edge weight $w_{ij} = \varphi\left(\frac{\|x_i - x_j\|_2}{h}\right)$, where $h > 0$ is a bandwidth parameter. We use $G_{N,h}$ to denote the random graph generated in this way.

Degree asymptotics Given a sequence of graphs $\{G_{N,h}\}_{N=1}^{\infty}$ generated as above, we study the behavior of the minimizers of J_p in the limit $N \to \infty$. A first step is to understand the behavior of the graph itself, and in particular its degree distribution in this limit. In order to gain intuition for this issue, consider the special case when $\varphi(z) = \mathbf{1}\{z \le 1\}$. If the bandwidth parameter $h > 0$ is held fixed, then any sequence $x_{i_1}, \cdots, x_{i_k}$ of points that fall in a ball of radius h will form a clique. Thus, the graph will contain roughly $1/h^d$ cliques, each with approximately Nh^d vertices. It is typically desired that the sequence of graphs be sparse with an appropriate degree growth (e.g., constant or logarithmic growth) so that it converges to the underlying manifold. In order to enforce this behavior, the bandwidth parameter h should tend to zero as the sample size N increases.

Under this assumption, it can be shown that the *scaled degree* at any vertex x, given by

$$d(x) = \frac{1}{Nh^d} \sum_{i=1}^{N} \varphi\left(\frac{\|x_i - x\|_2}{h}\right),$$

concentrates around $\mu(x)$. A precise statement can be found in Hein (2006a); roughly speaking, it follows from the fact that for any fixed point x in $[0,1]^d$, as h goes to zero and under a smoothness assumption on the density μ, the probability that a random vector $x_i \sim \mu$ falls in the h-neighborhood of x scales as $\Pr(\|x_i - x\|_2 \le h) = \int_{\|z-x\|_2 < h} \mu(z)dz \sim h^d \mu(x)$.

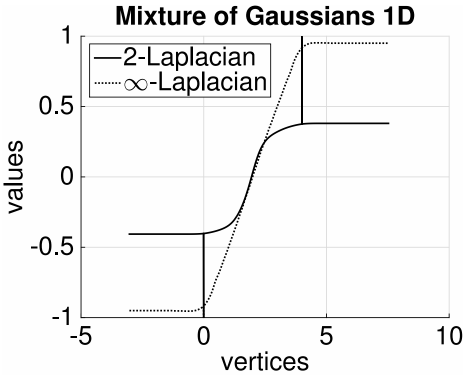


Figure 1: A mixture of two 1-dimensional Gaussians $N(0,1)$ and $N(4,1)$ with equal weights. 500 points are drawn i.i.d. from each component. We added one point at 0 with label -1 and one point at 4 with label +1. The similarity graph is constructed with an RBF kernel with bandwidth $.4$.

3. Variational problem and related PDE

In this section, our main goal is to study the behavior of the solution in the limit as the sample size $N \to \infty$ and the bandwidth $h \to 0$. As discussed above, it is natural to consider scalings under which h decreases in parallel with the increase in N. However, for simplicity, we follow Nadler et al. (2009) and first take the sample size N to infinity with the bandwidth held constant, and then let h go to zero.[2] Our first result characterizes the asymptotic behavior of the objective J_p.

Theorem 1 *Let f be continuously differentiable with a bounded derivative, and let μ be a bounded density. Then for any even integer $p \geq 2$, we have*

$$I_p(f) := \lim_{h\to 0} \lim_{N\to\infty} \frac{1}{N^2\, h^{p+d}} J_p(f) = C_p \int \|\nabla f(x)\|_2^p \mu^2(x) dx, \tag{6}$$

where $C_p := \frac{1}{d^{p/2}} \int \|z\|_2^p \varphi\big(\|z\|_2\big)^p dz$.

We provide the proof in Appendix A; it involves applying the strong law of large numbers for U-statistics, as well as an auxiliary result on the isotropic nature of the integral of a rank-one tensor.

We pause here to mention that the convergence of the objective function for a *fixed* function f does not imply convergence of the sequence of estimators $(f_{N,h})$ obtained by minimizing J_p. Obtaining a result of the latter form would require proving a statement of uniform convergence of J_p on a suitably large class of functions, and would take us far from the scope of the present paper. We point to Hein (2006b); Lu (2012) for such attempts in the case $p = 2$. Proving similar results for higher p bears a deeper study and is interesting in it's own right.

Based on Theorem 1, the asymptotic limit of the semi-supervised learning problem is a supervised non-parametric estimation problem with a regularization term given by the functional I_p—

[2] In the case $p = 2$, it is known that the same limiting objects are recovered when a joint limit in N and h is taken with $h \to 0$ but $Nh^d / \log N \to \infty$ (Hein (2006a)). Based on the discussion above, this scaling implies a super-logarithmic degree sequence. It is still to be verified if the same result holds for all p.

namely, the problem

$$\inf_g \int \|\nabla g(x)\|_2^p \, \mu^2(x) dx \quad \text{subject to } g(x_i) = y_i \text{ for all } i \in O. \tag{7}$$

Our next main result characterizes the solutions of this optimization problem in terms of a partial differential equation known as the (weighted) p-Laplacian equation. Here the word "weighted" refers to the term μ^2 in the functional (7) (Heinonen et al. (2012); Oberman (2013)).

Let us introduce various pieces of notation that are useful in the sequel. Given a smooth vector field $F : \mathbb{R}^d \to \mathbb{R}^d$, we use $\text{div}(F) := \sum_{i=1}^d \partial_{x_i} F_i$ to denote denote its divergence. For a scalar-valued function $f : \mathbb{R}^d \to \mathbb{R}$, we let

$$\Delta_2 f = \text{div}\left(\nabla f\right) = \sum_{i=1}^d \partial_{x_i}^2 f, \quad \text{and} \quad \Delta_\infty f = \frac{\langle \nabla f,\, \nabla^2 f \, \nabla f \rangle}{\langle \nabla f, \nabla f \rangle} = \frac{1}{\|\nabla f\|_2^2} \sum_{i,j=1}^d \partial_{x_i} f \cdot \partial_{x_i,x_j} f \cdot \partial_{x_j} f$$

denote the (standard) 2-Laplacian operator and the ∞-Laplacian operator, respectively.

Theorem 2 *Suppose that the density μ is bounded and continuously differentiable. Then any twice-differentiable minimizer f of the functional* (7) *must satisfy the Euler-Lagrange equation*

$$\text{div}\left(\mu^2(x)\|\nabla f(x)\|_2^{p-2} \nabla f(x)\right) = 0. \tag{8a}$$

If moreover the distribution μ has full support, then equation (8a) *is equivalent to*

$$\Delta_2 f(x) + 2\langle \nabla \log \mu(x), \nabla f(x) \rangle + (p-2)\Delta_\infty f(x) = 0. \tag{8b}$$

The proof employs standard tools from calculus of variations (Gelfand and Fomin, 1963). We note here that f does not need to be twice differentiable for the above result to hold (Heinonen et al. (2012)), in which case equations (8a) and (8b) have to be understood in the viscosity sense (Crandall et al. (2001); Armstrong and Smart (2010)). Twice differentiability is assumed so only for ease of the proof (see Appendix B).

When μ is the uniform distribution, equation (8b) reduces to the partial differential equation (PDE)

$$\Delta_2 f(x) + (p-2)\Delta_\infty f(x) = 0,$$

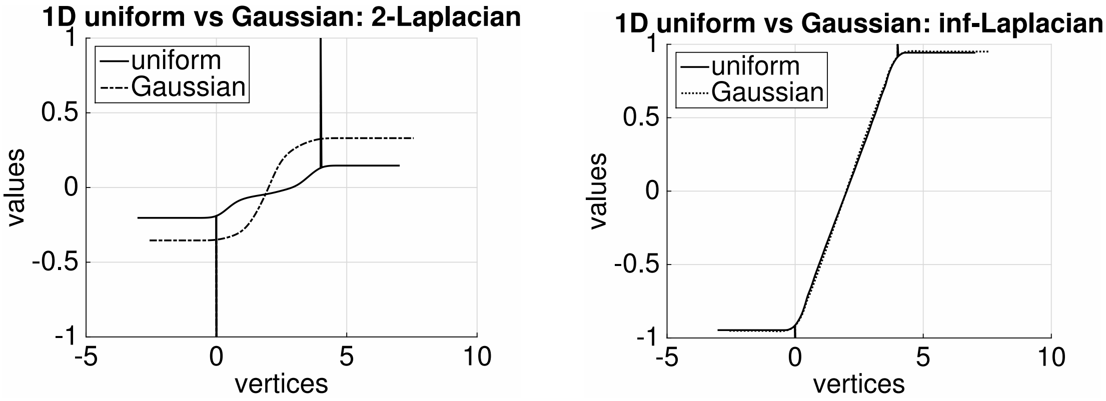
which is known as the p-Laplacian equation and often studied in the PDE literature (Heinonen et al. (2012); Oberman (2013)). If one divides by p and lets $p \to \infty$, one obtains the ∞-Laplacian equation

$$\Delta_\infty f(x) = 0, \tag{9}$$

subject to measurement constraints $f(x_i) = y_i$ for $i \in O$. This problem has been studied by various authors (e.g., Crandall et al. (2001); Aronsson et al. (2004); Peres et al. (2009); Armstrong and Smart (2010)). Note that in dimension $d = 1$, we have $\Delta_\infty = \Delta_2 = \frac{d^2}{dx^2}$, and equation (8b) reduces to

$$(p-1)\mu(x)f^{''}(x) + 2\mu'(x)f'(x) = 0.$$

Therefore, if we specialize to the case $p = 2$, the 2-Laplacian regularization method solves the differential equation $\mu(x)f^{''}(x) + 2\mu'(x)f'(x) = 0$, whereas if we specialize to $p = \infty$, then the inf-minimization method solves the differential equation $f^{''} = 0$. Note that the two equations coincide only when μ is the uniform distribution, in which case μ' is uniformly zero.



(a) (b)

Figure 2: Behavior of the 2- and infinity- Laplacian solutions under a change of input density. The latter is either a mixture of two 1-dimensional Gaussians $N(0,1)$ and $N(4,1)$ or a mixture of two uniform distributions $U([-3,3])$ and $U([1,7])$ with equal weights. In each case, 500 points are drawn i.i.d. from each component. The methods are given two observations $(0,-1)$ and $(4,1)$. (a) ℓ_2-based solution. (b) ℓ_∞-based solution.

4. Insights and predictions

Our theory from the previous section allows us to make a number of predictions about the behavior of different regularization methods, which we explore in this section.

4.1. Inf-minimization is insensitive to μ

Observe that the effect of the data-generating distribution μ has disappeared in equation (9). One could also see this by taking the limit $p \to \infty$ in the objective to be minimized in Theorem 1—in particular, assuming that μ has full support, we have $I_p(f)^{1/p} \to \sup_{[0,1]^d} \|\nabla f(x)\|_2$.

From the observations above, one can see that in the limit of infinite unlabeled data—i.e., once the distribution μ is available—the 2-Laplacian regularization method, as well as any p-Laplacian method based on finite (even) p, incorporates knowledge of μ in computing the solution; in contrast, for $p = \infty$, the inf-minimization method does not (see Figure 2). On the other hand, it has been shown that the 2-Laplacian method is badly behaved for $d \geq 2$ in the sense that the solution tends to be uninformative (constant) everywhere except on the points of observation. The solution must then have sharp variations on those points in order to respect the measurement constraints (see Figure 1 for an illustration of this phenomenon). We show in the next section that this problem plagues the p-Laplacian minimization approach as well whenever $p \leq d$.

4.2. p-Laplacian regularization is degenerate for $p \leq d$

In this section, we show that p-Laplacian regularization is degenerate for all $p \leq d$. This issue was originally addressed by Nadler et al. (2009), who provided an example that demonstrates the degeneracy for $p = 2$ and $d \geq 2$. Here we show that the underlying idea generalizes to all pairs

(p, d) with $p \leq d$. Recall that Theorem 1 guarantees that

$$I_p(f) := \lim_{h \to 0} \lim_{N \to \infty} \frac{1}{N^2 \, h^{p+d}} \, J_p(f) = C_p \int \|\nabla f(x)\|_2^p \mu^2(x) dx.$$

In the remainder of our analysis, we treat the cases $p \leq d-1$ and $p = d$ separately.

Case $p \leq d-1$: Beginning with the case $p \leq d-1$, we first set $x_0 = 0$ and then let x_1 be any point on the unit sphere (i.e., $\|x_1\|_2 = 1$). Define the function $f_\epsilon(x) = \min\{\|x\|_2/\epsilon, 1\}$ for some $\epsilon \in (0,1)$, and let the observed values be $y_j = f_\epsilon(x_j)$ for $j \in \{0,1\}$. Using the fact that $\nabla \|x\|_2 = \frac{x}{\|x\|_2}$ and assuming that μ is uniformly upper bounded by $\mu_{\max}$ on $[0,1]^d$, we have

$$I_p(f_\epsilon) = \int_{B(0,\epsilon)} \frac{\mu^2(x)}{\epsilon^p} dx \leq \frac{\mu_{\max}^2}{\epsilon^p} \operatorname{vol}(B(0,\epsilon)) = \mu_{\max}^2 \operatorname{vol}(B(0,1)) \epsilon^{d-p},$$

where $B(0,\epsilon)$ denotes the Euclidean ball of radius ϵ centered at the origin, and vol denotes the Lebesgue volume in $\mathbb{R}^d$. Consequently, we have $\lim_{\epsilon \to 0} I_p(f_\epsilon) = 0$, so the infimum of I_p is achieved for the trivial function that is 1 everywhere except at the origin, where it takes the value 0. The key issue here is that $\|\nabla f(x)\|_2^p$ grows at a rate of $\frac{1}{\epsilon^p}$ while the measure of the "spike" in the gradient shrinks at a rate of ϵ^d.

Case $p = d$: On the other hand, when $p = d$, we take $f_\epsilon(x) = \log\left(\frac{\|x\|_2^2 + \epsilon}{\epsilon}\right) / \log\left(\frac{1+\epsilon}{\epsilon}\right)$, for which we also have $y_0 = f_\epsilon(x_0) = 0$ and $y_1 = f_\epsilon(x_1) = 1$. With this choice, we have

$$\begin{aligned}
I_p(f_\epsilon) &= \frac{1}{\log\left(\frac{1+\epsilon}{\epsilon}\right)^d} \int_{B(0,1)} \frac{\|x\|_2^d}{\left(\|x\|_2^2 + \epsilon\right)^d} \mu^2(x) dx \\
&\leq \frac{\mu_{\max}^2}{\log\left(\frac{1+\epsilon}{\epsilon}\right)^d} \int_{B(0,1)} \frac{\|x\|_2^d}{\left(\|x\|_2^2 + \epsilon\right)^d} dx \\
&\overset{(i)}{=} \frac{\mu_{\max}^2}{\log\left(\frac{1+\epsilon}{\epsilon}\right)^d} \cdot \operatorname{vol}(B(0,1)) \int_0^1 \frac{r^d}{\left(r^2 + \epsilon\right)^d} d r^{d-1} dr \\
&\overset{(ii)}{=} \frac{d \, \mu_{\max}^2 \operatorname{vol}(B(0,1))}{\log\left(\frac{1+\epsilon}{\epsilon}\right)^d} \int_0^1 \frac{u^{d-1}}{\left(u + \epsilon\right)^d} du \\
&\overset{(iii)}{\leq} \frac{d \, \mu_{\max}^2 \operatorname{vol}(B(0,1))}{2 \log\left(\frac{1+\epsilon}{\epsilon}\right)^{d-1}},
\end{aligned}$$

where step (i) follows from a change of variables from x to the radial coordinate r; step (ii) follows by the variable change $u = r^2$; and step (iii) follows by upper-bounding u by $u + \epsilon$ in the numerator inside the integral. Again, we find that $\lim_{\epsilon \to 0} I_p(f_\epsilon) = 0$. Thus, in order to avoid degeneracies, it is necessary that $p \geq d+1$.

It is worth noting that Alamgir and Luxburg (2011) studied the problem of computing the so-called *q-resistances* of a graph, which are a family of distances on the graph having a formulation similar —in fact, dual— to the p-Laplacian regularization method considered in the present paper, and where $1/p + 1/q = 1$. They established a phase transition in the value of q for the geometric random graph model, where above the threshold $q^{**} = 1 + 1/(d-2)$, the q-resistances "[...] depend on trivial local quantities and do not convey any useful information [...]" about the graph, while below the threshold $q^* = 1 + 1/(d-1)$, these resistances encode interesting properties of the graph. They conclude by suggesting the use of p-Laplacian regularization with $1/p + 1/q^* = 1$. The latter condition can be read $p = d$. However, as shown by the examples above, this choice is still problematic, and in fact, the choice $d+1$ is the smallest admissible value for p.

We also note that the example for $p \leq d-1$ extends to an arbitrary number of labeled points: one simply has a spike for each point. Undesirable behavior arises as long as the set $\{x_i \mid i \in O\}$ of observed points is of measure zero. Finally, we note that both the above example can be adapted to the case where the squared loss $(f(x_i) - y_i)^2$ is optimized along with the regularizer instead of imposing the hard constraint $f(x_i) = y_i$ (see Appendix D). The issue is that the regularizer is too weak and allows to choose the solution from a very large class of functions.

4.3. p-Laplacian solution is smooth for $p \geq d+1$

At this point, a natural question is whether the condition $p \geq d+1$ is also *sufficient* to ensure that the solution is well-behaved. In a specific setting to be described here, the answer is *yes*[3]. The underlying reason is the Sobolev embedding theorem. More precisely, let $W^{1,p}([0,1]^d)$ denote the weighted Sobolev space of all (weakly) differentiable functions f on $[0,1]^d$ such that

$$\|f\|_{1,p} := \left(\int \|\nabla f(x)\|_2^p \mu^2(x) dx\right)^{1/p} < \infty.$$

The above is a semi-norm on $W^{1,p}([0,1]^d)$. If moreover, we assume μ is strictly positive almost everywhere and restrict the above class to functions vanishing on the boundary, then $\|\cdot\|_{1,p}$ actually defines a norm. When $p > d$, and under additional regularity conditions on μ (e.g. upper- and lower-bounded by constants a.e.), the space $W^{1,p}$ can be embedded continuously into the space of Hölder functions of exponent $1 - \frac{d}{p}$, i.e. functions u such that $|u(x) - u(y)| \leq c\|x-y\|_2^{1-\frac{d}{p}}$ for all $x, y \in [0,1]^d$ for some dimension-dependent constant c. For details, see Theorem 11.34 of Leoni (2009) or Lemma 5.17 of Adams and Fournier (2003). Brown and Opic (1992) provide some relaxed conditions on μ. Since the minimizer f of I_p is such that $I_p(f) = \int \|\nabla f(x)\|_2^p \mu^2(x) dx < \infty$, the function f is in the Sobolev class $W^{1,p}$, and therefore it automatically inherits the Hölder smoothness property, i.e. the p-Laplacian solution is smooth for $p \geq d+1$, asymptotically as $N \to \infty, h \to 0$[4]. Incidentally, via the examples in the previous section, it is clear that no such embedding exists if $p \leq d$.

4.4. An example where inf-minimization interpolates well

By extension to the case $p \to \infty$, the infinity-Laplacian solutions also enjoy continuity (this solution is actually Lipschitz based on its interpretation as the *absolutely minimal Lipschitz extension* of the observations (Aronsson et al. (2004)). It was also argued by Kyng et al. (2015) based on experimental results that the inf-minimization method has a better behavior in higher dimensions in terms of faithfulness to the observations. We illustrate this point by considering a simple example,

[3] Interestingly enough, the p-Laplacian equation has been extensively studied in non-linear potential theory. It is in fact the prototypical example of a non-linear degenerate elliptic equation. The regularity of the solutions is well understood for any real number $1 < p < \infty$ (see e.g. Heinonen et al. (2012)). For our purposes however, we do not need the full power of this theory.

[4] If the graph is finite, then the solution might still contain small spikes as apparent in Figures 1 and 2.

similar to the one above, for which the ∞-Laplacian equation (9) produces a sensible solution. With $x_0 = 0$ and $S := \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$ denoting the Euclidean unit sphere, suppose that we limit ourselves to functions satisfying the observation constraints $y_0 = f(0) = 0$ and $y(x) = 1$ for all $x \in S$.

Without any further information on the data-generating process, a reasonable fit is the function $\bar{f}(x) = \|x\|_2$. We claim that it is the only radially symmetric solution to the differential equation $\Delta_\infty f = 0$ with the boundary constraints $f(x_0) = y_0$ and $f(x) = y(x)$ for all $x \in S$. In order to verify this claim, let $f(x) = g(\|x\|_2)$ for a differentiable function $g : \mathbb{R}_+ \to \mathbb{R}$. For any non-zero $x \in \mathbb{R}^d$, we have

$$\nabla f(x) = g'(\|x\|_2)\frac{x}{\|x\|_2}, \quad \text{and} \quad \nabla^2 f(x) = g''(\|x\|_2)\frac{xx^\intercal}{\|x\|_2^2} + g'(\|x\|_2)\frac{I}{\|x\|_2} - g'(\|x\|_2)\frac{xx^\intercal}{\|x\|_2^3}.$$

Then

$$\Delta_\infty f(x) = \frac{1}{\|\nabla f(x)\|_2^2}\nabla f(x)^\intercal \nabla^2 f(x) \nabla f(x) = g''(\|x\|_2).$$

Given the boundary conditions on f, the only solution to $\Delta_\infty f$ = 0, is given by $g(r) = r$, meaning that $f(x) = \|x\|_2$. On the other hand, the latter is not a solution to $\Delta_2 f = 0$, unless $d = 1$.

In summary, this section reveals a trade-off between smoothness and sensitivity to the data-generating density μ in p-Laplacian regularization: the solution is strongly sensitive to μ but is non-smooth for small values of p, while it is smooth but weakly dependent on μ for large and infinite values of p. The transition from degeneracy to smoothness happens at a sharp threshold $p^* = d + 1$, while the dependence on μ weakens with larger and larger p without a threshold.

While the property of smoothness is an obvious quality in an estimation setting, and may lead to improved statistical rates if assumed first hand —especially if the signal one wishes to recover is itself smooth— it is less obvious how to quantify the advantages entailed by the sensitivity to the underlying data-generating density; especially when the latter is available and is to be incorporated in the design of an estimator. We provide in the next section a simple, one-dimensional regression example where the regression function is tied to the density μ via the cluster assumption, and where a difference in estimation rates between the ℓ_2 and ℓ_∞ methods is exhibited. This difference is explicitly due to the fact that the ℓ_2 method leverages the knowledge of μ while ℓ_∞ does not.

5. The price of "forgetting" μ

We consider in this section a simple estimation example in one dimension where the asymptotic formulation of 2-Laplacian regularization method achieves a better rate of convergence than that of the inf-minimization method. Such an advantage of the ℓ_2 method over the ℓ_∞ method should be conceivable under the cluster assumption: the regularizer $I_2 = \int f'^2\mu^2$ will encodes information about the target function via μ while the regularizer $I_\infty = \sup|f'|$ does not.

Let the target function f^* and the data-generating density μ be supported on the interval $[-1, 1]$. For some small $\epsilon > 0$, we construct a density μ that takes a uniformly small value over the interval $[-\epsilon, \epsilon]$, and takes and a large value on the complementary set $[-1, -\epsilon) \cup (\epsilon, 1]$. We also let f^* have a high Lipschitz constant on the interval $[-\epsilon, \epsilon]$ and be constant otherwise. More precisely, the density μ and function f^* are constructed as follows:

$$\mu(x) = \begin{cases} b & x \in [-\epsilon, \epsilon], \\ a & x \in [-1, -\epsilon) \cup (\epsilon, 1], \end{cases} \qquad f^*(x) = \begin{cases} -1 & x \in [-1, -\epsilon), \\ x/\epsilon & x \in [-\epsilon, \epsilon], \\ 1 & x \in (\epsilon, 1]. \end{cases} \tag{10}$$

The constants a and b are related by the equation $(1-\epsilon)a + \epsilon b = 1/2$ so that the density μ integrates to 1, and we think of b as being much smaller than a, i.e. $b \ll a$. Consider the following two classes of functions corresponding to the regularizer I_p for $p = 2$ and $p = \infty$ respectively:

$$\mathcal{H} := \Big\{ f : [-1,1] \to \mathbb{R} \ , \ f \text{ absolutely continuous, odd and } \int_{-1}^{1} f'(x)^2 \mu^2(x) dx < \infty \Big\},$$
$$\mathcal{L} := \Big\{ f : [-1,1] \to \mathbb{R} \ , \ f \text{ absolutely continuous, odd and } \sup_{|x| \le 1} |f'(x)| < \infty \Big\}.$$

We define the associated norms $\|f\|_{\mathcal{H}} := \big(\int_{-1}^{1} f'(x)^2 \mu^2(x) dx \big)^{1/2}$ and $\|f\|_{\mathcal{L}} := \sup_{x \in [-1,1]} |f'(x)|$ on $\mathcal{H}$ and $\mathcal{L}$ respectively. Observe that $\|f^*\|_{\mathcal{L}} = 1/\epsilon$ while $\|f^*\|_{\mathcal{H}}$ is upper bounded by a constant:

$$\int_{-1}^{1} [(f^*)'(x)]^2 \mu^2(x) dx = \int_{-\epsilon}^{\epsilon} \frac{1}{\epsilon^2} b^2 dx = 2b^2/\epsilon.$$

Taking $b = \sqrt{\epsilon}$, the above integral is bounded above by 2.

We draw n points $(x_i)_{i=1}^n$ independently from μ and observe the responses $y_i = f^*(x_i) + \sigma^2 \xi_i$ where $\xi_i \sim N(0,1)$ are i.i.d. standard normal random variables, $\sigma > 0$. We compare the following two M-estimators:

$$\widehat{f}_{\mathcal{H}} = \arg\min_f \frac{1}{n} \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 \quad \text{s.t. } f \in \mathcal{H} \ , \ \|f\|_{\mathcal{H}} \le 2, \qquad \text{and} \qquad \widehat{f}_{\mathcal{L}} = \arg\min_f \frac{1}{n} \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 \quad \text{s.t. } f \in \mathcal{L} \ , \ \|f\|_{\mathcal{L}} \le 1/\epsilon,$$

in terms of the rate of decay of the error $\big\| \widehat{f} - f^* \big\|_n^2$, where $\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^n f^2(x_i)$. Note that this error can in principle tend to zero as n grows to infinity since the target f^* belongs to the hypothesis class in both of the considered cases, i.e. there is no approximation error.

Theorem 3 *There are universal constants (c_0, c_1, c_2, c_3) such that for any $\epsilon \in (0, 1/2)$, the ℓ_2 estimator satisfies the bound*

$$\Big\| \hat{f}_{\mathcal{H}} - f^* \Big\|_n^2 \le c_1 \left(\frac{\sigma^2}{n} \right)^{2/3} \tag{11}$$

with probability at least $1 - \exp\big\{ -c_0 \big(\frac{n}{\sigma^2} \big)^{1/3} \big\}$. On the other hand, the ℓ_∞ estimator satisfies the bound

$$\big\| \widehat{f}_{\mathcal{L}} - f^* \big\|_n^2 \le c_3 \Big(\frac{\sigma^2}{\epsilon \, n} \Big)^{2/3} \tag{12}$$

with probability at least $1 - \exp\big\{ -c_2 \big(\epsilon^2 \frac{n}{\sigma^2} \big)^{1/3} \big\}$.

We can thus compare the upper bounds on the rate of estimation of the ℓ_2 and ℓ_∞ methods respectively. The upper bound (12) shows a dependence in $1/\epsilon^{2/3}$, while the bound (11) shows no dependence on ϵ. One might ask if these bounds are tight. In particular, a question of interest is whether the ℓ_∞ estimator *adapts* to the target f^* without the need to know the density μ, in which case the corresponding estimator would achieve a better rate. While we think our bounds could be sharpened, we strongly suspect that the ℓ_∞ estimator cannot achieve a rate independent of ϵ (in contrast to the ℓ_2 estimator). We provide an array of simulations showing that the rate of ℓ_∞ deteriorates as ϵ gets small, where the rate of the ℓ_2 method stays the same (see Figure 3).

5.1. Main ideas of the proof

The non-asymptotic bound (12) in Theorem 3 follows in a straightforward way from known results on the minimax rate of estimation on the class of Lipschitz functions. Indeed, the rate of estimation on this class with Lipschitz constant L is $\left(L\sigma^2/n\right)^{2/3}$, and in our case $L = 1/\epsilon$. On the other hand, the result (11) follows by recognizing that the class $\mathcal{H}$ is a weighted Sobolev space of order 1, which is a Reproducing Kernel Hilbert Space (RKHS). The associated kernel, as identified by Nadler et al. (2009), is given by

$$\mathcal{K}(x,y) = \frac{1}{4}\int_{-1}^{1} dt/\mu^2(t) - \frac{1}{2}\left|\int_{x}^{y} dt/\mu^2(t)\right|, \quad \text{for all } x, y \in [-1,1]. \tag{13}$$

It is known that the rate of estimation on a ball of radius R of an RKHS is tightly related to the decay of the eigenvalues of the kernel. More precisely, the rate of estimation is upper-bounded with high probability by the smallest solution $\delta > 0$ to the inequality

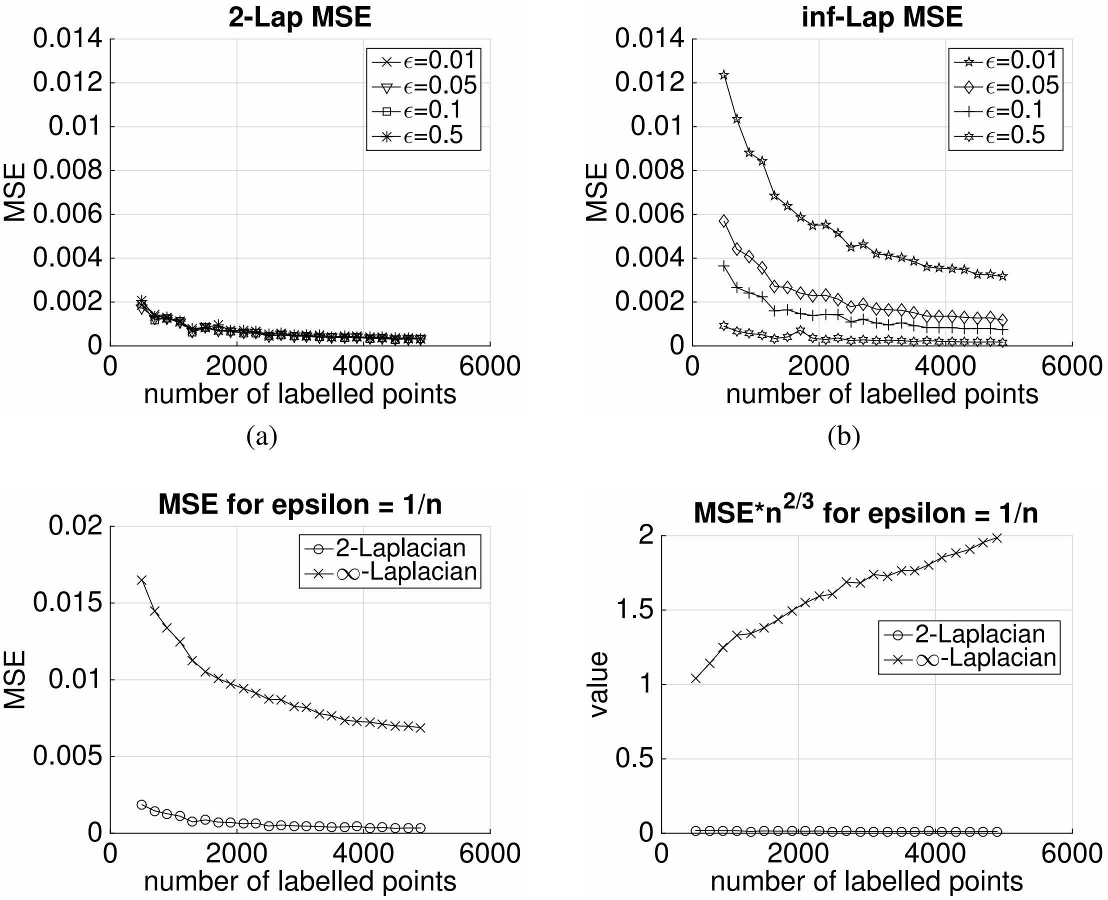
$$\left(\frac{2}{n}\sum_{j=0}^{\infty} \min\left\{\gamma_j, \delta^2\right\}\right)^{1/2} \leq \frac{R}{\sigma}\,\delta^2, \tag{14}$$

where $(\gamma_j)_{j\geq 0}$ is the sequence of eigenvalues of the kernel $\mathcal{K}$ (Koltchinskii (2006); Mendelson (2002); Bartlett et al. (2002); van de Geer (2000)). Our next result upper-bounds the rate of decay of these eigenvalues.

Lemma 4 *For any $\epsilon \in (0, 1/2)$, the eigenvalues of the kernel $\mathcal{K}$ form a doubly indexed sequence $(\gamma_{k,j})$ with $0 \leq k \leq 2k_0 - 1, j \geq 0$, $k_0 = \lfloor\sqrt{2}\epsilon^{-3/4}\rfloor$. This sequence satisfies the upper bound*

$$\gamma_{k,j} \leq \begin{cases} 1.26 & \textit{if } k = j = 0 \\ \frac{1}{\left(\frac{k}{2\sqrt{2}} + j\epsilon^{-3/4}\right)^2 \pi^2} & \textit{otherwise.} \end{cases}$$

Plugging these estimates in equation (14) leads to the rates we claim in Theorem 3. The full details are given in Appendix C.



(a)

(b)

(c)

(d)

Figure 3: Plots of mean-squared-error (MSE) against number of labeled samples. The regularization parameter is determined using cross-validation (thereby producing estimators of *superior* performance than $\widehat{f}_{\mathcal{H}}$ and $\widehat{f}_{\mathcal{L}}$). The samples x_i are drawn according to μ and $y_i = f^*(x_i) + \xi_i$, and ξ_i are i.i.d. $N(0, 0.05)$. Panels (a) and (b): MSE of ℓ_2 and ℓ_∞ methods respectively for various values of ϵ. As expected, the MSE of the ℓ_2 method is independent of ϵ while that of the ℓ_∞ method is increasing in $1/\epsilon$. Panels (c) and (d): Plots of the MSE and MSE $\times$ $n^{2/3}$, respectively versus the sample sample size n for both methods. Both plots correspond to sequences of problems with $\epsilon = 1/n$; panel (d) shows that in this regime, the rate of the ℓ_2 method is roughly $n^{-2/3}$, thereby providing evidence that the upper bound (11) is tight, while the rate of the ℓ_∞ method, call it $r(\epsilon, n)$, is such that $r(\frac{1}{n}, n) \gg n^{-2/3}$.

6. Related work

The discrete graph p-Laplacian was introduced by Zhou and Schölkopf (2005) as a form of regularization generalizing classical Laplacian regularization in semi-supervised learning Zhu et al. (2003). We mention however that the continuous p-Lapalcian has been extensively studied much earlier in PDE theory Heinonen et al. (2012); Aronsson et al. (2004). It was also used and analyzed for spectral clustering, where it provides a family of relaxations to the normalized cut problem Amghibech (2003); Bühler and Hein (2009); Luo et al. (2010). A dual version of the problem (*the q-resistances* or *q-volatges* problem) was investigated and shown to yield improved classification performance in Bridle and Zhu (2013). Alamgir and Luxburg (2011) prove the existence of a phase transition under the geometric random graph model roughly similar to the one exhibited in this paper, although the thresholds are slightly different. The exact nature of the connection is still unclear however. On the other hand, a game-theoretic interpretation of the p-Laplacian solution is studied in Peres and Sheffield (2008); Peres et al. (2009), and a similar transition at $p = d + 1$ in the behavior of the game is found. The assumption that the graph entails a geometric structure is popular in the analysis of semi-supervised learning algorithms Belkin and Niyogi (2001); Bousquet et al. (2003); Belkin and Niyogi (2004); Hein (2006a,b); Nadler et al. (2009). This line of work have mostly focused on the 2-Laplacian formulation and its convergence properties to a differential operator on the limiting manifold.

Among other approaches that circumvent the degeneracy issue discussed in the paper, we mention higher order regularization Zhou and Belkin (2011), where instead of only penalizing the first derivative of the function, one can penalize up to l derivatives. This approach considers solutions in a higher order Sobolev space $W^{l,2}$ which, via the Sobolev embedding theorem, only contains smooth functions if $l > d/2$ (see Adams and Fournier (2003); Leoni (2009)). This approach can be implemented algorithmically using the discrete *iterated Laplacian* Zhou and Belkin (2011); Wang et al. (2015).

Results on statistical rates for semi-supervised learning problems are very sparse. The first results are covered in Castelli and Cover (1996) in the context of mixture models, Rigollet (2007) in the context of classification, and Lafferty and Wasserman (2007) for regression. A recent line of work considers the setting where the graph is fixed while the set of vertices where the labels are available is random Ando and Zhang (2007); Johnson and Zhang (2007, 2008); Shivanna and Bhattacharyya (2014); Shivanna et al. (2015). The methods studied in this setting generalize the 2-Laplacian method by penalizing by the quadratic form given by a general positive semidefinite kernel. The derived rates depend on the structural properties of the graph and/or the kernel used. It is shown in particular that using the normalized Laplacian instead of the regular one leads to a better statistical bound, and on the other hand, one can obtain rates depending on the *Lovász Theta function* of the graph by choosing the regularization kernel optimally.

7. Conclusion and open problems

In this paper, we used techniques and ideas from PDE theory to yield insight into the behavior of various methods for semi-supervised learning on graphs. The d-dimensional geometric random graph model analyzed in this paper is a common one in the literature, and the most common Laplacian penalization technique is for $p = 2$, though the choice $p = \infty$ has also attracted some recent attention. Our paper sheds light on both of these options, as well as the full range of p in between. From our asymptotic analysis, we see that for a d-dimensional problem, degenerate solutions occur whenever $p \leq d$, whereas at the other extreme, the choice $p = \infty$ leads to solutions that are totally insensitive to the input data distribution. Hence, the choice $p = d + 1$ seems like a prudent one, trading off degeneracy of the solution with sensitivity to the unlabeled data.

An important companion problem is the unconstrained version of the problem, in which we penalize a (weighted) sum of two terms, the p-Laplacian term and a sum of squared losses on the labeled data. One can see that under our asymptotics, we can make the same conclusions about the unconstrained solution. Hence, the conclusions of this paper do not hinge on the fact that we modeled the problem with equality constraints, and do apply more generally. For completeness, we outline this argument in Appendix D.

In addition to the questions of obtaining uniform convergence results for the discrete p-Laplacian objective, and studying the convergence of the latter under a joint limit $N \to \infty$ and $h_N \to 0$ (i.e. the bandwidth parameter h decays to 0 as a function of N) already raised in the paper, there are perhaps two important assumptions which must be relaxed in future work. The first is the asymptotics we consider, in which the number of labeled points is fixed as the unlabeled points become infinite. An interesting situation is when both labeled and unlabeled points grow at a relative rate. We showed that in the first situation, a certain class of methods, namely all p-Laplacian methods with $p \leq d$, behave poorly. An interesting direction is to understand what set of methods are appropriate in different regimes of relative growth rate. In particular, we suspect that most of our results should continue to hold as long as the number of unlabeled points grow at a much faster rate than the number of labeled points.

The second assumption is about the geometric random graph model, and how much our results are tied to this model. Finite sample rates and non-asymptotic arguments are necessary to understand how soon we can expect to see these effects on general graphs in practice. The model selection problem of what p to use in practice is very important, since we may not know the underlying dimensionality of the data from which our graph was formed.

Finally, the question of designing fast algorithms for computing the estimator $\widehat{f}$ is of utmost importance. One promising idea is the use of a Newton method for minimizing the p-Laplacian objective: each iteration requires solving a Laplacian linear system, which is fast in theory (the Hessian matrix of J_p is the Laplacian of a graph with changing weights). This is the algorithm used in the experiments in the present paper, although we did not conduct a theoretical analysis of its overall running time.

Acknowledgments: We thank Jason Lee, Kevin Jamieson, Anup Rao, Sushant Sachdeva and Ryan Tibshirani for discussions in the early phases of this work. This work was partially supported by the Air Force Office of Scientific Research grant FA9550-14-1-0016 and the Office of Naval Research grant DOD-ONR-N00014 to MJW, the Office of Naval Research under grant number N00014-11-1-0688 and by the Army Research Laboratory and the U. S. Army Research Office under grant number W911NF-11-1-0391to MIJ.

References

Robert A Adams and John JF Fournier. *Sobolev spaces*, volume 140. Academic press, 2003.

Morteza Alamgir and Ulrike V Luxburg. Phase transition in the family of p-resistances. In *Advances in Neural Information Processing Systems*, pages 379–387, 2011.

S Amghibech. Eigenvalues of the discrete p-Laplacian for graphs. *Ars Combinatoria*, 67:283–302, 2003.

Rie Kubota Ando and Tong Zhang. Learning on graph with Laplacian regularization. *Advances in neural information processing systems*, 19:25, 2007.

Scott N Armstrong and Charles K Smart. An easy proof of Jensen's theorem on the uniqueness of infinity harmonic functions. *Calculus of Variations and Partial Differential Equations*, 37(3-4): 381–384, 2010.

Gunnar Aronsson, Michael Crandall, and Petri Juutinen. A tour of the theory of absolutely minimizing functions. *Bulletin of the American mathematical society*, 41(4):439–505, 2004.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized Rademacher complexities. In *Computational Learning Theory*, pages 44–58. Springer, 2002.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, volume 14, pages 585–591, 2001.

Mikhail Belkin and Partha Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine learning*, 56(1-3):209–239, 2004.

Olivier Bousquet, Olivier Chapelle, and Matthias Hein. Measure based regularization. In *Advances in Neural Information Processing Systems*, pages 1221–1228, 2003.

Nick Bridle and Xiaojin Zhu. p-voltages: Laplacian regularization for semi-supervised learning on high-dimensional data. In *Eleventh Workshop on Mining and Learning with Graphs (MLG2013)*, 2013.

RC Brown and B Opic. Embeddings of weighted sobolev spaces into spaces of continuous functions. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 439, pages 279–296. The Royal Society, 1992.

Thomas Bühler and Matthias Hein. Spectral clustering based on the graph p-Laplacian. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 81–88. ACM, 2009.

Vittori Castelli and Thomas M Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42 (6):2102–2117, 1996.

Michael G Crandall, Lawrence C Evans, and Ronald F Gariepy. Optimal Lipschitz extensions and the infinity Laplacian. *Calculus of Variations and Partial Differential Equations*, 13(2):123–139, 2001.

RB Darst, DA Legg, and DW Townsend. The Pólya algorithm in L_∞ approximation. *Journal of Approximation Theory*, 38(3):209–220, 1983.

R. M. Dudley. *Uniform central limit theorems*. Cambridge university press, 1999.

AG Egger and GD Taylor. Dependence on p of the best L_p approximation operator. *Journal of approximation theory*, 49(3):274–282, 1987.

Alan Egger and Robert Huotari. Rate of convergence of the discrete Pólya algorithm. *Journal of approximation theory*, 60(1):24–30, 1990.

IM Gelfand and SV Fomin. Calculus of variations. revised english edition translated and edited by richard a. silverman, 1963.

Matthias Hein. *Geometrical aspects of statistical learning theory*. PhD thesis, TU Darmstadt, 2006a.

Matthias Hein. Uniform convergence of adaptive graph-based regularization. In *Learning Theory*, pages 50–64. Springer, 2006b.

Juha Heinonen, Tero Kilpeläinen, and Olli Martio. *Nonlinear potential theory of degenerate elliptic equations*. Courier Corporation, 2012.

Rie Johnson and Tong Zhang. On the effectiveness of Laplacian normalization for graph semi-supervised learning. *Journal of Machine Learning Research*, 8(4), 2007.

Rie Johnson and Tong Zhang. Graph-based semi-supervised learning and spectral kernel design. *Information Theory, IEEE Transactions on*, 54(1):275–288, 2008.

Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Rasmus Kyng, Anup Rao, Sushant Sachdeva, and Daniel A Spielman. Algorithms for Lipschitz learning on graphs. *Proceedings of The 28th Conference on Learning Theory*, pages 1190–1223, 2015.

John Lafferty and Larry Wasserman. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems 20*, pages 801–808. Curran Associates, Inc., 2007.

David A Legg and Douglas W Townsend. The Pólya algorithm for convex approximation. *Journal of mathematical analysis and applications*, 141(2):431–441, 1989.

Giovanni Leoni. *A first course in Sobolev spaces*, volume 105. American Mathematical Society Providence, RI, 2009.

Yibiao Lu. Statistical methods with application to machine learning and artificial intelligence. 2012.

Dijun Luo, Heng Huang, Chris Ding, and Feiping Nie. On the eigenvectors of p-Laplacian. *Machine Learning*, 81(1):37–51, 2010.

Shahar Mendelson. Geometric parameters of kernel machines. In *Computational Learning Theory*, pages 29–43. Springer, 2002.

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Semi-supervised learning with the graph laplacian: The limit of infinite unlabelled data. In *Neural Information Processing Systems*, pages 1330–1338, 2009.

Adam M Oberman. Finite difference methods for the infinity Laplace and p-Laplace equations. *Journal of Computational and Applied Mathematics*, 254:65–80, 2013.

Yuval Peres and Scott Sheffield. Tug-of-war with noise: A game-theoretic view of the p-laplacian. *Duke Mathematical Journal*, 145(1):91–120, 2008.

Yuval Peres, Oded Schramm, Scott Sheffield, and David Wilson. Tug-of-war and the infinity Laplacian. *Journal of the American Mathematical Society*, 22(1):167–210, 2009.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8:2183–2206, 2007.

Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Rakesh Shivanna and Chiranjib Bhattacharyya. Learning on graphs using orthonormal representation is statistically consistent. In *Advances in Neural Information Processing Systems*, pages 3635–3643, 2014.

Rakesh Shivanna, Bibaswan K Chatterjee, Raman Sankaran, Chiranjib Bhattacharyya, and Francis Bach. Spectral norm regularization of orthonormal representations for graph transduction. In *Advances in Neural Information Processing Systems*, pages 2206–2214, 2015.

Sara van de Geer. *Empirical processes in M-estimation.* Cambridge University Press Cambridge, 2000.

Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan J Tibshirani. Trend filtering on graphs. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics,*, page 10421050, 2015.

Dengyong Zhou and Bernhard Schölkopf. Regularization on discrete spaces. In *Pattern Recognition*, pages 361–368. Springer, 2005.

Xueyuan Zhou and Mikhail Belkin. Semi-supervised learning by higher order regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 892–900, 2011.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of The 31st International Conference on Machine Learning*, volume 3, pages 912–919, 2003.

Appendix A. Proof of Theorem 1

Let x and x' be i.i.d. draws from μ. Since f is a bounded function, the first moment $\mathbb{E}\left[|f(x) - f(x')|^q\right]$ is finite. Therefore, by the strong law of large numbers for U-statistics (Serfling (2009)), we have

$$\lim_{N\to\infty} \frac{1}{N^2} J_p(f) = \int\int \varphi\left(\frac{\|x-x'\|_2}{h}\right)^p \left|f(x) - f(x')\right|^p \mu(x)\mu(x')dxdx', \quad \text{almost surely.}$$

Writing $x' = x + hz$ for some scalar $h > 0$ and vector z, the second integral simplifies to

$$\int \varphi\left(\frac{\|x-x'\|}{h}\right)^p \left|f(x) - f(x')\right|^p \mu(x')dx' = h^d \int \varphi(\|z\|_2)^p \left|f(x) - f(x+hz)\right|^p \mu(x+hz)dz.$$

We now divide by h^{d+p} and consider the behavior as the bandwidth parameter h tends to zero. Since the functions f, f' and μ are all bounded on a compact domain, the dominated convergence theorem implies that

$$\mu(x)\int \varphi\big(\|z\|_2\big)^p\,|\langle \nabla f(x), z\rangle|^p\,dz = \mu(x)\big\langle \nabla f(x)^{\otimes p}, \int \varphi\big(\|z\|_2\big)^p z^{\otimes p} dz\big\rangle.$$

Note the above inner product involves tensors of order p, and recall that we have assumed that $p \geq 2$ is even. Since the function φ depends only on the norm of z in the above integral, the latter should also be isotropic. The precise statement is as follows:

Lemma 5 *For any function $w : \mathbb{R}_+ \to \mathbb{R}_+$ and vector $u \in \mathbb{R}^d$, we have*

$$\big\langle u^{\otimes p}\ ,\ \int w\big(\|z\|_2\big) z^{\otimes p} dz\big\rangle = \begin{cases} \frac{1}{d^{p/2}}\big(\int w\big(\|z\|_2\big)\|z\|_2^p dz\big)\cdot \|u\|_2^p & \textit{if } p \geq 2 \textit{ is even} \\ 0 & \textit{if } p \textit{ is odd.} \end{cases} \tag{15}$$

Applying Lemma 5 with the function $w(\|z\|_2) := \varphi(\|z\|_2)^p$ and then simplifying yields

$$\lim_{N\to\infty} \frac{1}{N^2 h^{p+d}} J_p(f) = \frac{1}{d^{p/2}} \int \|z\|_2^p \varphi\big(\|z\|_2\big)^p dz \cdot \int \|\nabla f(x)\|_2^p \mu^2(x) dx,$$

which concludes the proof of the theorem.

The only remaining detail is to prove Lemma 5.

Proof of Lemma 5 We proceed by induction on the integer p. In the base case ($p = 2$), we have

$$\big\langle uu^\intercal\ ,\ \int w\big(\|z\|_2\big) zz^\intercal dz\big\rangle = u^\intercal\big(\int w\big(\|z\|_2\big) zz^\intercal dz\big)u.$$

Since the function w depends only on $\|z\|_2$, the matrix between the parentheses above is proportional to the identity, the proportionality constant can be determined by taking a trace. We end up with $\int w\big(\|z\|_2\big)zz^\intercal dz = \frac{1}{d}\big(\int w\big(\|z\|_2\big)\|z\|_2^2 dz\big)I$. This establishes the base case.[5] Now assume that for a given even $p \geq 2$, and for all function non-negative maps w, one has (15). We prove that the same is true for $p+2$. Define $T_p := \int w\big(\|z\|_2\big) z^{\otimes p} dz$, and for any vector $u \in \mathbb{R}^d$, let $\bar{T}_p$ be the partial contraction of T_{p+2} by $u \otimes u$, namely

$$\bar{T}_p := T_{p+2}(u \otimes u) = \int w\big(\|z\|_2\big) z^{\otimes p} \langle u, z\rangle^2 dz.$$

The tensor $\bar{T}$ is of order p, and the map $z \to w(\|z\|_2)\langle u, z\rangle^2$ is non-negative, so by the induction hypothesis, for every $v \in \mathbb{R}^d$, we have

$$\langle v^{\otimes p}\ ,\ \bar{T}_p\rangle = \frac{1}{d^{p/2}}\left(\int w\left(\|z\|_2\right)\|z\|_2^p \langle u, z\rangle^2 dz\right)\|v\|_2^p.$$

By recourse to the base case, the quadratic form between the parentheses is equal to $\frac{1}{d}\left(\int w\big(\|z\|_2\big)\|z\|_2^{p+2} dz\right)\|u\|_2^2$. Taking $u = v$ completes the proof of the lemma.

[5] The case $p = 2$ is also proven in Proposition 4.1 of the paper (Bousquet et al., 2003)

Appendix B. Proof of Theorem 2

Recall the shorthand notation $I_p(f) := \int \|\nabla f(x)\|_2^p \mu^2(x) dx$. By convexity, the function f is a minimizer of the functional I_p if for all test functions h and all sufficiently small real numbers $\epsilon > 0$, we have $I_p(f + \epsilon h) \geq I_p(f)$. Moreover, by a Taylor series expansion, we have

$$I_p(f + \epsilon h) = I_p(f) + p\epsilon \int \langle \nabla f(x), \nabla h(x) \rangle \cdot \|\nabla f(x)\|_2^{p-2} \mu^2(x) dx + \mathcal{O}(\epsilon^2),$$

where the $\mathcal{O}(\epsilon^2)$ term is non-negative by convexity of I_p. Hence, the function f is a minimizer if and only if

$$\int \langle \nabla f(x), \nabla h(x) \rangle \cdot \|\nabla f(x)\|_2^{p-2} \mu^2(x) dx = 0$$

for all testing functions h. By integrating by parts and choosing h to vanish on the boundary of the set $[0, 1]^d$, we find that the above quantity is equal to

$$\int \langle \nabla f(x), \nabla h(x) \rangle \cdot \|\nabla f(x)\|_2^{p-2} \mu^2(x) dx = - \int \operatorname{div} \big(\mu^2(x) \|\nabla f(x)\|_2^{p-2} \nabla f(x) \big) h(x) dx.$$

This expression has to vanish for all test functions h (that vanish on the boundary), which implies the Euler-Lagrange equation

$$\operatorname{div} \big(\mu^2(x) \|\nabla f(x)\|_2^{p-2} \nabla f(x) \big) = 0.$$

We now further manipulate this equation so as to obtain the p-Laplacian equation. In particular, some straightforward computations yield

$$\partial_{x_i} \big(\mu^2 \|\nabla f\|_2^{p-2} \partial_{x_i} f \big)(x) = \partial_{x_i} \big(\mu^2(x) \|\nabla f(x)\|_2^{p-2} \big) \partial_{x_i} f(x) + \mu^2(x) \|\nabla f(x)\|_2^{p-2} \partial_{x_i}^2 f(x), \quad \text{and}$$

$$\partial_{x_i} \big(\mu^2 \|\nabla f\|^{p-2} \big)(x) = 2\partial_{x_i} \mu(x) \cdot \mu(x) \|\nabla f(x)\|^{p-2} + \mu^2(x)(p-2) \big(\sum_{j=1}^d \partial_{x_i, x_j} f \partial_{x_j} f \big) \cdot \|\nabla f(x)\|^{p-4}.$$

Now summing these terms yield

$$\begin{aligned} \operatorname{div} \big(\mu^2(x) \|\nabla f(x)\|_2^{p-2} \nabla f(x) \big) &= 2\mu(x) \|\nabla f(x)\|_2^{p-2} \langle \nabla \mu(x), \nabla f(x) \rangle + \mu^2(x) \|\nabla f(x)\|_2^{p-2} \Delta_2 f(x) \\ &\quad + (p-2)\mu^2(x) \|\nabla f(x)\|_2^{p-4} \big(\sum_{i,j=1}^d \partial_{x_i} f \cdot \partial_{x_i, x_j} f \cdot \partial_{x_j} f \big)(x) \\ &= \mu^2(x) \|\nabla f(x)\|_2^{p-2} \cdot \big(\Delta_2 f(x) + \frac{2}{\mu(x)} \langle \nabla \mu(x), \nabla f(x) \rangle + (p-2) \Delta_\infty f(x) \big). \end{aligned}$$

From the derivation above, the Euler-Lagrange equation (8a) is equivalent to

$$\Delta_2 f(x) + 2 \langle \nabla \log \mu(x), \nabla f(x) \rangle + (p-2) \Delta_\infty f(x) = 0,$$

as claimed.

Appendix C. Proof of Theorem 3

Bounding the error of M-estimators is a classical problem in statistics and learning theory. Optimal rates typically follow by deriving uniform convergence bounds over a small ball *localized* around the true regression function f^*; for instance, see the book van de Geer (2000) as well as the papers Koltchinskii (2006); Bartlett et al. (2002). Uniform convergence is established by upper-bounding the Rademacher or Gaussian complexity of this small ball via generic covering number arguments, or by leveraging the special structure of the ball. The first approach will be used to analyze the rate of the estimator $\widehat{f}_{\mathcal{L}}$, and the second approach to analyze the rate of $\widehat{f}_{\mathcal{H}}$. In this latter case, the analysis is based on the study of the spectrum of a certain integral operator associated to the kernel $\mathcal{K}$ (13) that generates the space $\mathcal{H}$.

C.1. Proof of the error bound (12) on $\widehat{f}_{\mathcal{L}}$

For a given metric space $(\mathcal{F}, \rho)$, we let $N(t, \mathcal{F}, \rho)$ be the covering number of $\mathcal{F}$ in the metric ρ at resolution t. Now consider the shifted function class

$$\mathcal{L}^* := \{f - f^* \mid f \in \mathcal{L}\,,\ \|f\|_{\mathcal{L}} \leq 1/\epsilon\}$$

under the metric $\|f\|_{\mathcal{L}} = \sup_x |f'(x)|$. By known results on metric entropy (Dudley, 1999), we have $\log(N(t; \mathcal{L}^*; \|\cdot\|_{\mathcal{L}})) = \mathcal{O}\big(\frac{1}{\epsilon t}\big)$, using the fact that any function in $\mathcal{L}^*$ must be $2/\epsilon$-Lipschitz.

Now let $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f^2(x_i)$ be the squared empirical L^2-norm, and consider the ball

$$\mathbb{B}_n(\delta, \mathcal{L}^*) := \{f \in \mathcal{L}^* \mid \|f\|_n < \delta\}.$$

Since the sup norm is stronger than this empirical norm, we have the sequence of inequalities

$$\log(N(t; \mathbb{B}_n(\delta; \mathcal{L}^*); \|\cdot\|_n)) \leq \log(N(t; \mathcal{L}^*; \|\cdot\|_n)) \leq \log(N(t; \mathcal{L}^*; \|\cdot\|_{\mathcal{L}})) = \mathcal{O}\left(\frac{1}{\epsilon t}\right). \tag{16}$$

Next consider the δ-localized Gaussian complexity of a function class $\mathcal{F}$, given by

$$\mathbb{G}(\delta; \mathcal{F}) = \mathbb{E}_w\left[\sup_{\substack{g \in \mathcal{F} \\ \|g\|_n \leq \delta}} \frac{1}{n}\sum_{i=1}^n w_i g(x_i)\right], \tag{17}$$

where $\{w_i\}_{i=1}^N$ is an i.i.d. sequence of standard normal random variables. Define the critical radius δ_n as the smallest δ that satisfies the *master inequality*

$$\mathbb{G}(\delta; \mathcal{L}^*) \leq \frac{\delta^2}{2\sigma}. \tag{18}$$

With this set-up, it is known (van de Geer, 2000) that the M-estimator $\widehat{f}_{\mathcal{L}}$ satisfies a bound of the form

$$\mathbb{P}\left[\|\widehat{f}_{\mathcal{L}} - f^*\|_n^2 \leq c_1\delta_n^2\right] \geq 1 - e^{-c_2\frac{n\delta_n^2}{2\sigma^2}}, \tag{19}$$

where c_1 and c_2 are universal positive constants. By Dudley's entropy integral, the critical radius δ_n is upper bounded by any δ which satisfies

$$\frac{1}{\sqrt{n}} \int_{\delta^2/2}^{\delta} \sqrt{\log(N(t; \mathbb{B}_n(\delta; \mathcal{L}^*); \|\cdot\|_n))} dt \leq \frac{\delta^2}{\sigma}.$$

A little calculation shows that it suffices to choose δ_n such that

$$\delta_n^2 \leq \left(\frac{\sigma^2}{\epsilon n}\right)^{2/3}.$$

Note that this is in fact a global upper bound on δ_n, since the first inequality of (16) holds for any setting of the design points $\{x_i\}_{i=1}^n$.

C.2. Proof of the error bound (11) on $\widehat{f}_{\mathcal{H}}$

As our starting point, we use the master inequality (18) with the shifted function class $\mathcal{L}^*$ replaced by $\mathcal{H}^* := \{f - f^* \mid f \in \mathcal{H} \,,\, \|f\|_{\mathcal{H}} \leq R\}$. Recall that $\|f\|_{\mathcal{H}}^2 = \int f'^2 \mu^2$ and $R = 2$ in our case. Using known bounds (Mendelson, 2002) on localized Gaussian complexity of $\mathcal{H}^*$ in terms of the eigenvalues of the kernel $\mathcal{K}$, the master inequality takes the simpler form

$$\left(\frac{2}{n} \sum_{k,j=0}^{\infty} \min\left\{\gamma_{k,j}, \delta^2\right\}\right)^{1/2} \leq \frac{R}{\sigma}\, \delta^2. \tag{20}$$

We claim that this master inequality is satisfied by

$$\delta_n = c_1 \left(\frac{\sigma^2}{R^2 n}\right)^{1/3} \tag{21}$$

where c_1 is an absolute constant, independent of n, ϵ. Given this choice, the overall claim follows by applying the non-asymptotic error bound (19).

It remains to show that the choice (21) is valid, and we do so by using the bounds on the eigenvalues from Lemma 4. For $\delta \in (0, \sqrt{2}/\pi)$, let k^*, j^* be the largest integers such that $\gamma_{k,j} \geq \delta^2$, or equivalently such that

$$k/(2\sqrt{2}) + j\epsilon^{-3/4} \leq 1/(\pi\delta).$$

One can verify that $j^* = \lfloor \frac{\epsilon^{3/4}}{\delta\pi} \rfloor$ and $k^* = \lfloor 2\sqrt{2}\epsilon^{-3/4}(\frac{\epsilon^{3/4}}{\delta\pi} - \lfloor \frac{\epsilon^{3/4}}{\delta\pi} \rfloor) \rfloor$. We note that for $\delta < \sqrt{2}/\pi < 1$, j^* and k^* cannot simultaneously be zero. Using these cut-off points, the number of eigenvalues $(\gamma_{k,j})$ that are larger than δ^2 is at most $2k_0 j^* + k^* + 1$. Moreover, we have

$$\sum_{k,j=0}^{\infty} \min\left\{\gamma_{k,j}, \delta^2\right\} \leq (2k_0 j^* + k^* + 1)\delta^2 + \sum_{k=k^*+1}^{2k_0-1} \gamma_{k,j^*} + \sum_{j \geq j^*+1} \sum_{k=0}^{2k_0-1} \gamma_{k,j}. \tag{22}$$

Next, we control each term in the above expression. For the first term, note that $k_0 \leq \sqrt{2}\epsilon^{-3/4}$ and $j^* \leq \frac{\epsilon^{3/4}}{\delta\pi}$. Moreover, we have $k^* \leq \frac{2\sqrt{2}}{\pi\delta}$. Therefore, the first term is bounded as $2k_0 j^* + k^* + 1 \leq 4\sqrt{2}/(\delta\pi) + 1$.

As for the second term,

$$\begin{aligned}\sum_{k=k^*+1}^{2k_0-1}\gamma_{k,j^*} &\leq \sum_{k=k^*+1}^{2k_0-1}\frac{1}{\left(\frac{k}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)^2\pi^2}\\ &= \frac{1}{\left(\frac{k^*+1}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)^2\pi^2}+\sum_{k=k^*+2}^{2k_0-1}\frac{1}{\left(\frac{k}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)^2\pi^2}\\ &\overset{(i)}{\leq} \frac{1}{\left(\frac{k^*+1}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)^2\pi^2}+\int_{k^*+1}^{2k_0-1}\frac{dt}{\left(\frac{t}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)^2\pi^2}\\ &= \frac{1}{\left(\frac{k^*+1}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)^2\pi^2}+\frac{2\sqrt{2}}{\left(\frac{k^*+1}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)\pi^2}-\frac{2\sqrt{2}}{\left(\frac{2k_0-1}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)\pi^2}\\ &\leq \frac{1}{\left(\frac{k^*+1}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)^2\pi^2}+\frac{2\sqrt{2}}{\left(\frac{k^*+1}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)\pi^2}\\ &\overset{(ii)}{\leq} \frac{4\sqrt{2}}{\left(\frac{k^*+1}{2\sqrt{2}}+j^*\epsilon^{-3/4}\right)\pi^2}.\end{aligned}$$

The first inequality is obtained using Lemma 4. Inequality (i) follows by upper-bounding the discrete sum by an integral: $\sum_{k=k^*+1}^{k_0} 1/k^2 \leq \int_{k^*}^{k_0} dt/t$ (this argument will be used once more to control of the third sum). Inequality (ii) is obtained simply by factoring and noting that $\frac{k^*+1}{2\sqrt{2}} + j^*\epsilon^{-3/4} \geq \frac{1}{2\sqrt{2}}$. By definition of j^* and k^*, we have $\frac{k^*+1}{2\sqrt{2}} + j^*\epsilon^{-3/4} > \frac{1}{\pi\delta}$. Therefore,

$$\sum_{k=k^*+1}^{2k_0-1}\gamma_{k,j^*} \leq \frac{4\sqrt{2}}{\left(\frac{1}{\delta\pi}\right)\pi^2} = \frac{4\sqrt{2}}{\pi}\delta.$$

On the other hand, using the same techniques above, the third term is upper bounded as

$$\begin{aligned}\sum_{j\geq j^*+1}\sum_{k=0}^{2k_0-1}\gamma_{k,j} &\leq \sum_{j\geq j^*+1}\frac{2k_0}{(j\epsilon^{-3/4})^2\pi^2}\\ &= \frac{2k_0\epsilon^{3/2}}{(j^*+1)^2\pi^2}+\sum_{j\geq j^*+2}\frac{2k_0\epsilon^{3/2}}{j^2\pi^2}\\ &\leq \frac{2k_0\epsilon^{3/2}}{(j^*+1)^2\pi^2}+\int_{j^*+1}^{\infty}\frac{2k_0\epsilon^{3/2}}{s^2\pi^2}ds\\ &= \frac{2k_0\epsilon^{3/2}}{(j^*+1)^2\pi^2}+\frac{2k_0\epsilon^{3/2}}{(j^*+1)\pi^2}\\ &\leq 2\frac{2k_0\epsilon^{3/2}}{(j^*+1)\pi^2}.\end{aligned}$$

Since $j^* = \lfloor\frac{\epsilon^{3/4}}{\delta\pi}\rfloor$, we have $j^*+1 > \frac{\epsilon^{3/4}}{\delta\pi}$. Therefore, using $k_0 = \lfloor\sqrt{2}\epsilon^{-3/4}\rfloor$, we have

$$\sum_{j\geq j^*+1}\sum_{k=0}^{2k_0-1}\gamma_{k,j} \leq \frac{4k_0\epsilon^{3/4}}{\pi}\delta \leq \frac{4\sqrt{2}}{\pi}\delta.$$

Putting these estimates together, inequality (22) yields

$$\sum_{k,j=0}^{\infty} \min\left\{\gamma_{k,j}, \delta^2\right\} \le (\frac{4\sqrt{2}}{\delta\pi}+1)\delta^2 + \frac{8\sqrt{2}}{\pi}\delta \le \delta^2 + \frac{12\sqrt{2}}{\pi}\delta.$$

Therefore, the master inequality (20) becomes

$$\sqrt{\frac{2}{n}}\Big(\delta^2 + \frac{12\sqrt{2}}{\pi}\delta\Big)^{1/2} \le \frac{R}{\sigma}\delta^2.$$

Finally, only considering solutions $\delta < 1$, it can be verified that the specified choice (21) is adequate, as claimed.

C.3. Proof of Lemma 4

We first characterize the eigenvalues of the kernel $\mathcal{K}$ as solutions to a certain non-linear equation:

Lemma 6 *The eigenvalues of the kernel $\mathcal{K}$ from equation (13) are given by the solutions $\lambda > 0$ of the non-linear equation*

$$\tan\left(\frac{\epsilon}{\sqrt{\lambda b}}\right)\tan\left(\frac{1-\epsilon}{\sqrt{\lambda a}}\right) = \left(\frac{b}{a}\right)^{3/2}. \tag{23}$$

The proof of this lemma is deferred to the next subsection. Our next step is to understand the solutions to $\tan\left(\frac{\epsilon}{\sqrt{b}}x\right)\tan\left(\frac{1-\epsilon}{\sqrt{a}}x\right) = \left(\frac{b}{a}\right)^{3/2}$ with $x = 1/\sqrt{\lambda}$.

For our purposes, we only need to consider the regime where $b = \sqrt{\epsilon} \ll a$ and the quantity ϵ is close to zero. We then assume that the ratio of the two periods $\frac{1-\epsilon}{\sqrt{a}}/\frac{\epsilon}{\sqrt{b}} := k_0$ is an integer; note that this assumption can always be satisfied by choosing ϵ appropriately. This assumption prevents complicated oscillatory phenomena, and thus makes the analysis simpler.

Define the function $\phi(x) := \tan\left(\frac{\epsilon}{\sqrt{b}}x\right)\tan\left(\frac{1-\epsilon}{\sqrt{a}}x\right)$. We first exploit the periodicity of the function ϕ so as to simplify the reduce the problem ϕ is even, and by our assumption on k_0, it has a period of $\frac{\sqrt{a}}{1-\epsilon}\pi$. Therefore, we only study its behavior on $[0, \frac{\sqrt{b}}{\epsilon}\pi]$. We divide this interval into two intervals $I = [0, \frac{\sqrt{b}}{\epsilon}\pi/2)$ and $\bar{I} = [\frac{\sqrt{b}}{\epsilon}\pi/2, \frac{\sqrt{b}}{\epsilon}\pi]$ and we study the behavior of ϕ on each of them separately.

Divide I into smaller subintervals $I_k := \left[k\frac{\sqrt{a}}{1-\epsilon}\pi/2, (k+1)\frac{\sqrt{a}}{1-\epsilon}\pi/2\right)$ with $0 \le k \le k_0 - 1$. On each interval I_k, both functions $x \to \tan\left(\frac{1-\epsilon}{\sqrt{a}}x\right)$ and $x \to \tan\left(\frac{\epsilon}{\sqrt{b}}x\right)$ are continuous, positive, and increasing. In addition, the last one varies from 0 to ∞. Therefore, ϕ spans the entire half line $[0, \infty)$ on I_k. Consequently, it must cross the line $y = \left(\frac{b}{a}\right)^{3/2}$ exactly once in each interval I_k. We denote the coordinate of intersection by x_k, i.e. $\phi(x_k) = \left(\frac{b}{a}\right)^{3/2}$ and $x_k \in I_k$.

Similarly, we divide $\bar{I}$ into regular subintervals $\bar{I}_k = \frac{\sqrt{b}}{\epsilon}\pi - I_k$. We observe that by parity of ϕ, we have $\phi(-x_k) = \left(\frac{b}{a}\right)^{3/2}$, and by periodicity, $\bar{x}_k = -x_k + \frac{\sqrt{b}}{\epsilon}\pi \in \bar{I}_k$ also verifies $\phi(\bar{x}_k) = \left(\frac{b}{a}\right)^{3/2}$. The sequence of numbers $x_0 < x_1 < \ldots < x_{k_0-1} < \bar{x}_{k_0-1} < \bar{x}_{k_0-2} < \ldots < \bar{x}_0$ correspond to the entire set of solutions on the interval $[0, \frac{\sqrt{b}}{\epsilon}\pi]$. Then by periodicity, we obtain all positive solutions by translating the above sequence by multiples of the period $\frac{\sqrt{b}}{\epsilon}\pi$. Therefore the eigenvalues of the kernel $\mathcal{K}$ form a doubly-indexed sequence $(\gamma_{k,j})$ such that

$$\gamma_{k,j} = \begin{cases} 1\Big/\big(x_k + j\frac{\sqrt{b}}{\epsilon}\pi\big)^2 & k \in [0, k_0 - 1], \\ 1\Big/\big(\frac{\sqrt{b}}{\epsilon}\pi - x_{2k_0-k-1} + j\frac{\sqrt{b}}{\epsilon}\pi\big)^2 & k \in [k_0, 2k_0 - 1]. \end{cases}$$

Now, we can upper bound $\gamma_{k,j}$ by using the fact $x_k \in I_k$ when either k or j is greater than 1: observe that when $k \le k_0 - 1$, $x_k \ge k\frac{\sqrt{a}}{1-\epsilon}\pi/2$. Likewise, when $k_0 \le k \le 2k_0 - 1$,

$$\frac{\sqrt{b}}{\epsilon}\pi - x_{2k_0-k-1} \ge \frac{\sqrt{b}}{\epsilon}\pi - ((2k_0 - k - 1) + 1)\frac{\sqrt{a}}{1-\epsilon}\pi/2 = k\frac{\sqrt{a}}{1-\epsilon}\pi/2.$$

Thus, we have shown that $\gamma_{k,j} \le 1\Big/\big(k\frac{\sqrt{a}}{1-\epsilon}/2 + j\frac{\sqrt{b}}{\epsilon}\big)^2\pi^2$ for all $1 \le k \le 2k_0 - 1$. Furthermore, recalling that $a = \frac{1/2-\epsilon^{3/2}}{1-\epsilon}$, we notice that $\frac{\sqrt{a}}{1-\epsilon} \ge 1/\sqrt{2}$ when $\epsilon < 1/2$. Consequently, we have

$$\gamma_{k,j} \le 1\Big/\Big(\frac{k}{2\sqrt{2}} + j\frac{\sqrt{b}}{\epsilon}\Big)^2\pi^2 \qquad \text{whenever } k \ge 1 \text{ or } j \ge 1.$$

Note in passing that $k_0 = \frac{1-\epsilon}{\sqrt{a}}\epsilon^{-3/4} \le \sqrt{2}\epsilon^{-3/4}$.

The case $k = j = 0$ needs extra care. We have $x_0 \in [0, \frac{\sqrt{a}}{1-\epsilon}\pi/2)$. Since ϕ vanishes at 0 and is strictly increasing on this interval, it is clear that $x_0 > 0$. Now we proceed by an approximation argument valid in the limit $\epsilon \to 0$, and then invoke monotonicity of the solution x_0 in ϵ. For ϵ sufficiently small and by using $b = \sqrt{\epsilon}$, we can uniformly approximate the function $\tan\big(\frac{\epsilon}{\sqrt{b}}x\big)$ by $\frac{\epsilon}{\sqrt{b}}x$ for $0 < x < 1$. Then, the equation $\phi(x) = \big(\frac{b}{a}\big)^{3/2}$ becomes $x\tan\big(\frac{1-\epsilon}{\sqrt{a}}x\big) = \frac{\sqrt{b}}{\epsilon}\big(\frac{b}{a}\big)^{3/2} = \frac{1}{a^{3/2}}$. In this regime $a \simeq 1/2$ and the equation can be further approximated by $\tan(\sqrt{2}x) = \frac{2\sqrt{2}}{x}$. A numerical inspection shows that the latter has a unique solution $.892 \le x^* \le .899$ on $[0, 1]$. Moreover we also numerically observe that the solution $x_0(\epsilon)$ to the equation $\phi(x) = \big(\frac{b}{a}\big)^{3/2}$ on $[0, 1]$ is an increasing function of ϵ, hence $x_0 \ge \lim_{\epsilon\to 0} x_0(\epsilon) = x^*$. Therefore, the first eigenvalue $\gamma_{0,0} = 1/x_0^2$ is upper bounded by a constant independent of ϵ—that is, we have $\gamma_{0,0} \le 1/x^{*2} \le 1.26$, as claimed.

C.4. Proof of Lemma 6

Letting P be the distribution associated with the density μ, we study the eigenvalues and eigenfunctions of the integral operator $T : L_2(P) \to L_2(P)$ given by

$$Tf(x) := \int_{-1}^{1} f(t)\mathcal{K}(t, x)\mu(t)dt.$$

Here the reader should recall that the density μ takes the form

$$\mu(x) = \begin{cases} b & \text{if } x \in [-\epsilon, \epsilon] \\ a & \text{if } x \in [-1, -\epsilon) \cup (\epsilon, 1]. \end{cases}$$

where the parameters (a, b, ϵ) are related by the equations $(1-\epsilon)a + \epsilon b = 1/2$, and $b = \sqrt{\epsilon}$. The kernel $\mathcal{K}$ is given by

$$\mathcal{K}(x,y) = \frac{1}{4}\int_{-1}^{1} \frac{dt}{\mu^2(t)} - \frac{1}{2}\left|\int_{x}^{y} \frac{dt}{\mu^2(t)}\right|, \quad \text{for } x, y \in [-1, 1].$$

The eigenvalue equation associated with the operator T can be written as $T\varphi_\lambda = \lambda\varphi_\lambda$, where φ_λ is the eigenfunction associated with the eigenvalue $\lambda \geq 0$. Differentiating this equation twice yields a system of differential equations for the eigenfunctions (proof omitted):

Lemma 7 *All eigenfunctions of T must satisfy the following system of differential equations:*

$$\begin{aligned} &\lambda b\, \varphi_\lambda'' + \varphi_\lambda = 0 \ \text{ on } \quad [-\epsilon, \epsilon], \quad \textit{and} \\ &\lambda a\, \varphi_\lambda'' + \varphi_\lambda = 0 \quad \text{on} \quad [-1, -\epsilon) \ \cup \ (\epsilon, 1]. \end{aligned}$$

Solving this system of differential equations yields that any eigenfunction must be of the form

$$\varphi_\lambda(x) = \begin{cases} A_1 \sin\left(\frac{x}{\sqrt{\lambda b}}\right) & \text{for } x \in [-\epsilon, \epsilon], \\ A_2 \sin\left(\frac{x}{\sqrt{\lambda a}}\right) + B_2 \cos\left(\frac{x}{\sqrt{\lambda a}}\right) & \text{for } x \in [-1, -\epsilon), \text{ and} \\ A_2 \sin\left(\frac{x}{\sqrt{\lambda a}}\right) - B_2 \cos\left(\frac{x}{\sqrt{\lambda a}}\right) & \text{for } x \in (\epsilon, 1], \end{cases}$$

where we already exploited the fact that φ_λ has to be odd. Of course, not all functions of the above form are eigenfunctions of T, since we lost information by taking two derivatives. In order to show that φ_λ is actually an eigenfunction, we need to verify that it is continuous continuous, and satisfies the relations $\big(T\varphi_\lambda\big)' = \lambda\varphi_\lambda'$ and $T\varphi_\lambda = \lambda\varphi$. Actually, the last condition will be satisfied when the first two are. Together, these conditions will provide enough constraints to specify the four parameters A_1, A_2, B_2 and λ in an unambiguous, modulo a global multiplicative constant for the first three.

By imposing continuity on the solutions for $\pm\epsilon$, we obtain an equation relating the parameters of the problem:

$$A_2 \sin\left(\frac{\epsilon}{\sqrt{\lambda a}}\right) - B_2 \cos\left(\frac{\epsilon}{\sqrt{\lambda a}}\right) = A_1 \sin\left(\frac{\epsilon}{\sqrt{\lambda b}}\right). \tag{24}$$

Next we verify that the condition $\big(T\varphi_\lambda\big)' = \lambda\varphi_\lambda'$ holds on $[-\epsilon, \epsilon]$ and $[-1, -\epsilon) \cup (\epsilon, 1]$ separately. Since the density μ is even, we have $\mathcal{K}(-x, y) = \mathcal{K}(x, -y)$ for all $x, y \in [-1, 1]$. Therefore, for any odd function f, we have

$$\int_{-1}^{-\epsilon} f(t)\mathcal{K}(t,x)\mu(t)dt = -\int_{\epsilon}^{1} f(t)\mathcal{K}(t,-x)\mu(t)dt,$$

so one we can study the problem on the interval $(\epsilon, 1]$ and automatically obtain the corresponding results on $[-1, -\epsilon)$.

Case $x \in (\epsilon, 1]$: For any odd function f, we have

$$Tf(x) = -\frac{1}{2}\int_{-1}^{-\epsilon} af(t)\big(\frac{-\epsilon - t}{a^2} + \frac{2\epsilon}{b^2} + \frac{x-\epsilon}{a^2}\big)dt - \frac{1}{2}\int_{-\epsilon}^{\epsilon} bf(t)\big(\frac{\epsilon - t}{b^2} + \frac{x-t}{a^2}\big)dt \\ - \frac{1}{2}\int_{\epsilon}^{1} af(t)\big(\frac{|t-x|}{a^2}\big)dt.$$

Differentiating with respect to x and setting $f = \varphi_\lambda$ yields

$$\big(T\varphi_\lambda\big)'(x) = -\frac{1}{2a}\big\{A_2\big[-\sqrt{\lambda a}\cos\big(\frac{t}{\sqrt{\lambda a}}\big)\big]_{-1}^{-\epsilon} + B_2\big[\sqrt{\lambda a}\sin\big(\frac{t}{\sqrt{\lambda a}}\big)\big]_{-1}^{-\epsilon}\big\} \\ - \frac{1}{2a}\big\{A_2\big[-\sqrt{\lambda a}\cos\big(\frac{t}{\sqrt{\lambda a}}\big)\big]_{\epsilon}^{x} - B_2\big[\sqrt{\lambda a}\sin\big(\frac{t}{\sqrt{\lambda a}}\big)\big]_{\epsilon}^{x}\big\} \\ + \frac{1}{2a}\big\{A_2\big[-\sqrt{\lambda a}\cos\big(\frac{t}{\sqrt{\lambda a}}\big)\big]_{x}^{1} - B_2\big[\sqrt{\lambda a}\sin\big(\frac{t}{\sqrt{\lambda a}}\big)\big]_{x}^{1}\big\}.$$

Since we must have $\big(T\varphi_\lambda\big)' = \lambda\varphi_\lambda$, some algebra then leads to

$$A_2\cos\big(\frac{1}{\sqrt{\lambda a}}\big) + B_2\sin\big(\frac{1}{\sqrt{\lambda a}}\big) = 0. \tag{25}$$

Equations (24) and (25) form a linear system in the coefficients A_2 and B_2, and solving this system yields

$$A_2 = A_1\frac{\sin\big(\frac{1}{\sqrt{\lambda a}}\big)\sin\big(\frac{\epsilon}{\sqrt{\lambda b}}\big)}{\sin\big(\frac{1}{\sqrt{\lambda a}}\big)\sin\big(\frac{\epsilon}{\sqrt{\lambda a}}\big) + \cos\big(\frac{1}{\sqrt{\lambda a}}\big)\cos\big(\frac{\epsilon}{\sqrt{\lambda a}}\big)},$$
$$B_2 = -A_1\frac{\cos\big(\frac{1}{\sqrt{\lambda a}}\big)\sin\big(\frac{\epsilon}{\sqrt{\lambda b}}\big)}{\sin\big(\frac{1}{\sqrt{\lambda a}}\big)\sin\big(\frac{\epsilon}{\sqrt{\lambda a}}\big) + \cos\big(\frac{1}{\sqrt{\lambda a}}\big)\cos\big(\frac{\epsilon}{\sqrt{\lambda a}}\big)}.$$

Case $x \in [-\epsilon, \epsilon]$: For an odd function f, we have

$$Tf(x) = -\frac{1}{2}\int_{-1}^{-\epsilon} af(t)\big(\frac{-\epsilon - t}{a^2} + \frac{x+\epsilon}{b^2}\big)dt - \frac{1}{2}\int_{-\epsilon}^{\epsilon} bf(t)\big(\frac{|x-t|}{b^2}\big)dt \\ - \frac{1}{2}\int_{\epsilon}^{1} af(t)\big(\frac{\epsilon - x}{b^2} + \frac{t-\epsilon}{a^2}\big)dt.$$

Following an argument similar to the previous case, we find that

$$\big(T\varphi_\lambda\big)'(x) = -\frac{a}{b^2}\big\{A_2\big[-\sqrt{\lambda a}\cos\big(\frac{t}{\sqrt{\lambda a}}\big)\big]_{-1}^{-\epsilon} + B_2\big[-\sqrt{\lambda a}\sin\big(\frac{t}{\sqrt{\lambda a}}\big)\big]_{-1}^{-\epsilon}\big\} \\ - \frac{1}{2b}\big\{A_1\big[-\sqrt{\lambda b}\cos\big(\frac{t}{\sqrt{\lambda b}}\big)\big]_{-\epsilon}^{x} + A_1\big[-\sqrt{\lambda b}\cos\big(\frac{t}{\sqrt{\lambda b}}\big)\big]_{x}^{\epsilon}\big\}.$$

Imposing the constraint $\big(T\varphi_\lambda\big)' = \lambda\varphi_\lambda$ leads to the equation

$$A_2\big(\cos\big(\frac{\epsilon}{\sqrt{\lambda a}}\big) - \cos\big(\frac{1}{\sqrt{\lambda a}}\big)\big) + B_2\big(\sin\big(\frac{\epsilon}{\sqrt{\lambda a}}\big) - \sin\big(\frac{1}{\sqrt{\lambda a}}\big)\big) = \big(\frac{b}{a}\big)^{3/2}A_1\cos\big(\frac{\epsilon}{\sqrt{\lambda b}}\big). \tag{26}$$

Finally, plugging the expressions of A_2 and B_2 in the above equation and simplifying yields the claimed equation (23).

Appendix D. Reduction of least squares to constrained optimization

In this appendix, we show that in the regime $p \leq d$, minimizing an objective function that is a weighted sum of a least-squares cost with a regularization term $\left(\int \|\nabla f(x)\|^p \mu^2(x) dx\right)^{1/p}$ will have degenerate solutions, just like the constrained formulation. Consider a least-squares problem of the form

$$\min_f \Big\{ \sum_{i \in O} (f(x_i) - y_i)^2 + \lambda R(f) \Big\}, \tag{27}$$

where $R(f)$ is some arbitrary regularization term. We show that the above has the same solution as a constrained optimization problem. Let

$$\widehat{f} = \arg\min_f \Big\{ \sum_{i \in O} (f(x_i) - y_i)^2 + \lambda R(f) \Big\}.$$

Then $\widehat{f}$ is equal to the minimizer of

$$\min_f R(f) \quad \text{subject to} \;\; f(x_i) = \widehat{f}(x_i), i \in O. \tag{28}$$

To see this, suppose that the optimizer of (28) is $g \neq \widehat{f}$, then it must be that $R(g) < R(\widehat{f})$ and $\sum_{i \in O}(g(x_i) - y_i)^2 = \sum_{i \in O}(\widehat{f}(x_i) - y_i)^2$, in which case the function g achieves a smaller value of the cost (27) than $\widehat{f}$.

By setting $R(f) = \left(\int \|\nabla f(x)\|^p \mu^2(x) dx\right)^{1/p}$, we know from Section 4.2 that the solution to the optimization problem (28) must be degenerate, and so must the solution to the problem (27).