# Laplace Meets Moreau: Smooth Approximation to Infimal Convolutions Using Laplace's Method

Ryan J. Tibshirani[a] Samy Wu Fung[b] Howard Heaton[c] Stanley Osher[d]

[a]University of California, Berkeley [b]Colorado School of Mines [c]Typal Academy
[d]University of California, Los Angeles

**Abstract**

We study approximations to the Moreau envelope—and infimal convolutions more broadly—based on Laplace's method, a classical tool in analysis which ties certain integrals to suprema of their integrands. We believe the connection between Laplace's method and infimal convolutions is generally deserving of more attention in the study of optimization and partial differential equations, since it bears numerous potentially important applications, from proximal-type algorithms to Hamilton-Jacobi equations.

## 1 Introduction

Infimal convolutions are of core importance in mathematical optimization and partial differential equations (PDEs). The most well-known special case of an infimal convolution is the Moreau envelope, due to Moreau (1962, 1965), which (along with its counterpart, the proximal operator) is a key tool in convex and variational analysis, and in numerical algorithms for optimization. More broadly, beyond the Moreau envelope, infimal convolutions appear as solutions in a class of Hamilton-Jacobi equations in PDEs.

Laplace's method, due to Laplace (1774), is a tool for approximating integrals that finds applications in many areas of mathematics, statistics, physics, and computer science. More specifically, it provides a way to precisely approximate an integral whose integrand becomes increasingly peaked around its maximum value. To researchers in statistics and machine learning, Laplace's method is perhaps most familiar from its use in Bayesian inference, where it leads to an approximation of the posterior distribution in terms of a Gaussian distribution centered at the maximum a posteriori (MAP) estimate.

These two ideas are actually closely connected: Laplace's method provides a natural way to approximate an infimal convolution. This has been noted and used (albeit somewhat indirectly) by some authors in the past; see Section 1.4. However, we believe the connection between infimal convolutions and Laplace's method is not as widely appreciated as it should be, especially as it relates to sampling, which is a way to view (and numerically approximate) the integrals that appear in Laplace's method. Thus, the current paper places the connection between infimal convolutions and Laplace's method front and center. It is not really our intent to claim novelty in developing or formalizing this connection. Instead, our goal is to highlight both its elegance and utility in the hope that it gains better recognition, and potentially, sees further applications.

**Contributions.** To briefly highlight our contributions, we show that some recently proposed techniques for approximating Moreau envelopes and proximal operators, which have been motivated through a connection to PDEs, can be instead derived directly via self-normalized Laplace approximation. This allows us to extend the approximation technique to a broader class of problems, of infimal convolution form. We derive theory on the asymptotic validity of this approximation, which requires weaker conditions than the traditional analysis of Laplace approximation. We also present several example applications and numerical experiments.

**Outline.** In what follows, we first review preliminary concepts and related work. In Section 2, we describe the use of Laplace's method to approximate infimal convolutions, and we give interpretations from various perspectives. In Section 3, we derive approximation guarantees. In Section 4 we cover sampling techniques, and in Section 5, we walk through applications in optimization and PDEs, with illustrative examples.


Figure 1: Illustration of infimal convolution as epigraph addition (2) (here $g = \|\cdot\|_2^2/2$).

### 1.1 Infimal convolution

Let $f, g : \mathbb{R}^d \to \mathbb{R}$ be arbitrary real-valued functions. The *infimal convolution* (or simply the inf convolution) of $f$ and $g$ is another function, denoted $f \,\#\, g : \mathbb{R}^d \to \mathbb{R}$, which is defined by

$$(f \,\#\, g)(x) = \inf_y \Big\{ f(y) + g(x-y) \Big\}. \tag{1}$$

The notion of an infimal convolution originated with Fenchel (1951), while a related idea was independently developed over a series of papers by Bellman and Karush (1961, 1962a,b, 1963). Influenced by Fenchel, the papers by Moreau (1963) and Rockafellar (1963) serve as the basis for what is now considered the modern definition and treatment of inf convolutions, with Moreau's work providing the name "inf convolution", as well. For more on infimal convolutions, we refer to Moreau (1970); Strömberg (1994); Rockafellar and Wets (2009). The latter book introduces the notation # for the inf convolution operator that we use in this paper, which is meant to remind the reader of the addition operator, because (as Rockafellar and Wets emphasize) infimal convolution acts as addition on the space of epigraphs:

$$\mathrm{epi}(f \,\#\, g) = \mathrm{epi}(f) + \mathrm{epi}(g), \tag{2}$$

as long as the infimum defining $(f \,\#\, g)(x)$ is attained whenever finite. Here $\mathrm{epi}(f) = \{(x,t) : f(x) \le t\}$ is the epigraph of $f$, similarly for $\mathrm{epi}(g)$, and $A + B = \{a + b : a \in A,\, b \in B\}$. Figure 1 gives an illustration.

A particularly important special case of an inf convolution is the Moreau envelope. This plays a central role in various aspects of optimization, from theoretical to practical, and is covered in the next subsection. Inf convolutions also play an important role as solutions to certain Hamilton-Jacobi equations. To elaborate, let $H : \mathbb{R}^d \to \mathbb{R}$ be a convex Hamiltonian, and consider the first-order PDE

$$\begin{aligned} \partial_t u + H(\nabla u) &= 0, \qquad t > 0, \\ u(x,0) &= f(x), \quad t = 0, \end{aligned} \tag{3}$$

where $\partial_t u$ denotes the derivative of $u$ with respect to $t$, and $\nabla u$ denotes its gradient with respect to $x$. By the Hopf-Lax formula, (e.g., Theorem 6 in Chapter 3.3.2 of Evans (2010)), the solution is

$$u(x,t) = \inf_y \Big\{ f(y) + tH^*\Big(\frac{x-y}{t}\Big) \Big\}, \quad t > 0, \tag{4}$$

where $H^*$ is the conjugate (also called the Legendre-Fenchel transform) of $H$. In other words, the solution to the Hamilton-Jacobi PDE (3) at time $t > 0$ is given by the infimal convolution $u(x,t) = (f \,\#\, tH^*(\cdot/t))(x)$.

We finish this subsection with a useful general fact about infimal convolutions. Fix any point $x$, assume that there is a unique point $y_x$ which attains the infimum in (1), and assume $g$ is differentiable on $\mathbb{R}^d$. Then under some additional regularity conditions (e.g., Theorem 10.13 and Corollary 10.14 of Rockafellar and Wets (2009)), the inf convolution $f \,\#\, g$ is differentiable at $x$ and

$$\nabla(f \,\#\, g)(x) = \nabla g(x - y_x). \tag{5}$$

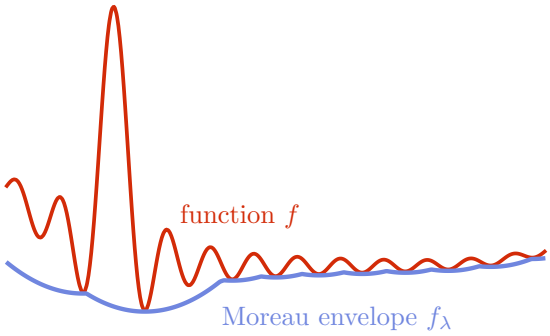



Figure 2: Illustration of the Moreau envelope $f_\lambda = f \,\#\, \|\cdot\|_2^2/(2\lambda)$ (for a particular $\lambda$).

## 1.2 Moreau regularization

For $g = \|\cdot\|_2^2/(2\lambda)$, with $\lambda > 0$ a fixed constant, the infimal convolution $f \,\#\, g$ is called the *Moreau envelope* of $f$ at the level $\lambda$, named after the pioneering work of Moreau (1962, 1965), and is denoted by

$$f_\lambda(x) = \inf_y \Big\{ f(y) + \frac{1}{2\lambda}\|x-y\|_2^2 \Big\}. \tag{6}$$

Figure 2 gives an illustration. Intimately connected to this is the *proximal operator* of $\lambda f$, denoted by

$$\mathrm{prox}_{\lambda f}(x) = \operatorname*{argmin}_y \Big\{ f(y) + \frac{1}{2\lambda}\|x-y\|_2^2 \Big\}, \tag{7}$$

where we use $\operatorname{argmin}_y F(y)$ to denote the set of minimizers of a function $F$, reducing to $\operatorname{argmin}_y F(y) = z$ if the set of minimizers is a singleton $\{z\}$. Importantly, if $f$ is convex, then the operator $\mathrm{prox}_{\lambda f}$ is guaranteed to be single-valued (rather than set-valued): it maps each input $x$ to a unique point $\mathrm{prox}_{\lambda f}(x)$.

The Moreau envelope (6) and its associated proximal operator (7) are ubiquitous throughout convex and variational analysis, as well as optimization. In terms of theory, proximal operators admit various important connections to subdifferentials, conjugates, and monotone operators; see, e.g., Rockafellar and Wets (2009); Bauschke and Combettes (2011). In terms of algorithms, proximal operators serve as a building block for a number of operator splitting techniques for nonsmooth, constrained, large-scale optimization, which includes forward-backward splitting, Douglas-Rachford splitting, and the alternating direction method of multipliers (ADMM); see, e.g., Boyd et al. (2011); Combettes and Pesquet (2011); Parikh and Boyd (2013); Beck (2017); Ryu and Yin (2022).

An early and influential contribution on the algorithmic front is called the *proximal point algorithm*, due to Rockafellar (1976). Given a function $f$ to be minimized, we fix $\lambda > 0$, initalize $x_0 \in \mathbb{R}^d$, and then repeat the iterations, for $k = 1, 2, 3, \ldots$:

$$x_k = \mathrm{prox}_{\lambda f}(x_{k-1}). \tag{8}$$

Returning to the gradient formula for $f \,\#\, g$ in (5), for a closed convex function $f$ (and with $g = \|\cdot\|_2^2/(2\lambda)$, which together satisfy the regularity conditions needed for this gradient formula), we have

$$(\nabla f_\lambda)(x) = \frac{x - \mathrm{prox}_{\lambda f}(x)}{\lambda}. \tag{9}$$

Interestingly, this is true regardless of the smoothness of $f$. Applying (9) to (8), we see that Rockafellar's proximal point iteration can be rewritten as

$$x_k = x_{k-1} - \lambda(\nabla f_\lambda)(x_{k-1}). \tag{10}$$

In other words, the proximal point algorithm is the same as gradient descent on the Moreau envelope, with step size equal to the envelope level $\lambda$.

## 1.3 Laplace's method

Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and $h : \mathbb{R}^d \to \mathbb{R}$ be continuous. *Laplace's method* (also called Laplace approximation) provides an asymptotic equivalence for an integral that becomes increasingly peaked around the global minimizer $x^\star$ of $\varphi$, assumed to be unique:

$$\int h(x) \exp(-t\varphi(x)) \, \mathrm{d}x \;\sim\; \frac{(2\pi/t)^{d/2}}{\exp(t\varphi(x^\star))} \frac{h(x^\star)}{\sqrt{\det(\nabla^2 \varphi(x^\star))}}, \tag{11}$$

where we use $a(t) \sim b(t)$ as shorthand for $a(t)/b(t) \to 1$ as $t \to \infty$.

In short, applications of Laplace's method can be found throughout statistics, probability, and machine learning (not to mention its uses in mathematics and physics). Examples include Bayesian computation and inference (e.g., Kass et al. (1991)), higher-order asymptotics (e.g., Shun and McCullagh (1995)), mean-field theory (where Laplace's method is often called the *saddle point approximation*, e.g., Mézard and Montanari (2009), Chapter 2), Gaussian processes (e.g., Rasmussen and Williams (2006), Chapter 3), and Bayesian deep learning (e.g., Daxberger et al. (2021)).

Now we present a formal statement of the validity of Laplace's approximation (11). We allow $\varphi$ to have an arbitrary domain $\mathcal{K}$, and reparametrize by $t = 1/\delta$ (taking $\delta \to 0^+$ in the asymptotic limit), since it will be more convenient for our purposes later.

**Theorem 1.** *Let $\varphi : \mathcal{K} \to \mathbb{R}$ be continuous over a compact set $\mathcal{K} \subseteq \mathbb{R}^d$. Assume that $\varphi$ has a unique global minimizer $x^\star$ in the interior of $\mathcal{K}$, and $\varphi$ is twice continuously differentiable on a neighborhood of $x^\star$, with strictly positive definite Hessian $\nabla^2 \varphi(x^\star)$. Then for any continuous function $h : \mathcal{K} \to \mathbb{R}$,*

$$\sqrt{\det(\nabla^2 \varphi(x^\star))} \; \frac{\exp(\varphi(x^\star)/\delta)}{(2\pi\delta)^{d/2}} \int_{\mathcal{K}} h(x) \exp(-\varphi(x)/\delta) \, \mathrm{d}x \;\to\; h(x^\star), \quad \text{as } \delta \to 0^+. \tag{12}$$

*This conclusion extends to the case where $\mathcal{K}$ is not compact (e.g., $\mathcal{K} = \mathbb{R}^d$) provided there exists $\epsilon > 0$ such that the sublevel set $\mathcal{S}_\epsilon = \{x \in \mathcal{K} : \varphi(x) \le \varphi(x^\star) + \epsilon\}$ is bounded and $\int_{\mathcal{K}} |h(x)| \exp(-d\varphi(x)/(2\epsilon)) \, \mathrm{d}x < \infty$.*

Proofs of the asymptotic convergence of Laplace's method can be found in any standard reference on the topic. For completeness, we provide a proof of Theorem 1 in Appendix A.1, based on the simple and elegant arguments given in Bach (2021).

The form of the approximation given in Theorem 1 (or equivalently in (11)) is well-suited for traditional applications of Laplace's method. In such problems, we seek to avoid computing an integral in $\exp(-t\varphi)$, the left-hand side in (11), and approximate it using a minimum of $\varphi$ (or maximum of $-\varphi$), the right-hand side in (11). For example, in Bayesian inference, this method can be used to approximate the posterior distribution by a Gaussian whose mean is the maximum a posteriori (MAP) estimate. Here, the sample size $n$ plays the role of $t$ in (11), making the approximation more accurate for larger sample sizes.

For our purposes however, a different form of the Laplace approximation will be more convenient. This is because our motivation is really the *opposite* of the traditional one: we seek to avoid computing a minimum of $\varphi$, and we instead approximate it using an integral involving $\exp(-t\varphi)$. In our setting, we would not want the approximation to feature normalizing constants such as $\exp(t\varphi(x^\star))$ and $\sqrt{\det(\nabla^2 \varphi(x^\star))}$, since they are unknown (recall, $\varphi(x^\star)$ is the minimum that we are trying to approximate in the first place). Fortunately, we can use the following *self-normalized* version of Laplace's method:

$$\frac{\int_{\mathcal{K}} h(x) \exp(-t\varphi(x)) \, \mathrm{d}x}{\int_{\mathcal{K}} \exp(-t\varphi(x)) \, \mathrm{d}x} \;\sim\; h(x^\star). \tag{13}$$

This follows directly from (12), by using the latter to approximate separately the integrals in the numerator and denominator in (13) (the common factor $(2\pi/t)^{d/2}/(\exp(t\varphi(x^\star))\sqrt{\det(\nabla^2 \varphi(x^\star))})$ cancels out). Because this just relies on two applications of Laplace's method, Theorem 1 gives the formal conditions under which (13) is valid. Later in Section 3, we will give a generalization of this result: instead of requiring $\varphi$ to be twice differentiable on a neighborhood of its minimizer, we only require it to be locally Hölder continous.

## 1.4 Related work

The literature related to the topic of our paper is vast, and below is our attempt to give a broad overview of the various related areas, without giving a comprehensive review of any one in particular.

**Viscosity solutions in PDEs.** In a certain sense, the use of Laplace's method in order to approximate solutions in Hamilton-Jacobi equations dates back to seminal work on *viscosity solutions* by Crandall and Lions (1983); Crandall et al. (1984) (see also Evans (1980)). A canonical way to construct viscosity solutions in Hamilton-Jacobi equations is the called the *vanishing viscosity* method, where we solve a modified PDE that has an additional diffusion term with a small viscosity parameter, and then send the viscosity parameter to zero. Interestingly, as we review in Section 2.2, for the Hamilton-Jacobi equation (3) where $H = \|\cdot\|_2^2/2$ (also called Burgers' equation), the use of a Laplacian diffusion term leads to exactly the same approximation as Laplace's method applied to the original solution. As far as we understand, some but not all authors in the PDE literature on viscosity make the connection to Laplace's approximation explicit. For example, it is not discussed in the early papers in the 1980s by Crandall, Lions, and Evans, but it can be found in Evans (2010) in Chapter 4.5.2. We have not yet seen the use of Laplace's approximation for general Hamilton-Jacobi equations (3) (for general $H$, as we propose in this paper), and as we discuss at the end of Section 2.2, we are unsure as to whether there is a viscosity-like representation for such an approximation in general.

In the total variation-based image denoising literature, the posterior mean formula (17), a special case of the Laplace approximation (16) for infimal convolutions that we consider in this paper (discussed in the next section), has been studied by Louchet (2008); Louchet and Moisan (2013). This was extended by Darbon and Langlois (2021); Darbon et al. (2021), who derive rigorous approximations guarantees using connections to Hamilton-Jacobi PDEs. Another line of work that draws connections between viscous Hamilton-Jacobi PDEs, proximal operators, and Moreau envelopes was initiated by Chaudhari et al. (2018) and further developed by Heaton et al. (2023); Osher et al. (2023). As we will explain shortly, in Sections 2.1 and 2.2, in this paper we arrive at the identical approximation for the proximal operator as that given in Darbon et al. (2021); Heaton et al. (2023); Osher et al. (2023), albeit from a different perspective: by directly applying Laplace's method, instead of relying on viscosity. By casting the approximation through the lens of Laplace's method, we are able to seamlessly extend it to handle arbitrary infimal convolutions. We are also able to make less stringent assumptions (on the functions in question) for the approximation theory that we derive in Section 3.

**Sampling and optimization.** As we explore in Sections 5.2 and 5.3, Laplace's approximation (17) of the proximal operator (and (16) for inf convolutions more generally) leads to various sampling-based methods for optimization. These methods are *zeroth-order*: they depend only on function evaluations (and not gradients, as would be the case in first-order methods). The connections between sampling and optimization are quite deep (in both directions—using sampling to optimize, and using optimization to sample). Classical examples include stochastic gradient descent (Robbins and Monro, 1951), simulated annealing (Kirkpatrick et al., 1983), and Langevin dynamics (Welling and Teh, 2011). More recently, stochastic localization and diffusion models have taken a center stage in machine learning (e.g., Sohl-Dickstein et al. (2015); Song and Ermon (2019); Ho et al. (2020); Song et al. (2021); El Alaoui et al. (2022); Montanari (2023)), and remain an extremely active topic of research.

In the setting of zeroth-order optimization in particular, the connections to sampling are also rich, dating back to Matyas (1965). Related to the Laplace approximation of the proximal point algorithm (studied in Section 5.2) is the idea of Gaussian smoothing from Nesterov and Spokoiny (2017), who study a gradient-free optimization algorithm which approximates a directional derivative with a finite difference scheme based on a random pertubation of the parameter. The motivation and focus in their work, as with much of the literature in zeroth-order optimization, is quite different than ours. We return to this discussion in Section 5.2.

## 2 Smooth approximation

Let $f, g : \mathbb{R}^d \to \mathbb{R}$ be continuous functions. Define for any fixed $x \in \mathbb{R}^d$ and $\delta > 0$:

$$y_x^\delta = \frac{\int y \exp\left(\frac{-f(y)-g(x-y)}{\delta}\right) \mathrm{d}y}{\int \exp\left(\frac{-f(y)-g(x-y)}{\delta}\right) \mathrm{d}y}. \tag{14}$$

Notice that each coordinate $(y_x^\delta)_i$ in (14) is the self-normalized Laplace approximation from (13), applied to $\varphi_x(y) = f(y) + g(x-y)$, with $h(y) = y_i$. In other words, we can view $y_x^\delta$ as approximating the minimizer of

the inf convolution criterion at $x$,

$$y_x^\delta \approx \operatorname*{argmin}_y \big\{ f(y) + g(x-y) \big\}.$$

This approximation becomes exact as $\delta \to 0^+$ under mild conditions on $f, g$, as the theory in the next section will make precise (see Corollary 1). It is important to note that convexity of $f, g$ is not required. Below, we discuss interpretations of the approximation (14), from different perspectives.

## 2.1 Exponential tilting

Define a density by

$$p_{g,x}^\delta(y) = \frac{\exp(-g(x-y)/\delta)}{\int \exp(-g(x-y)/\delta)\,\mathrm{d}y}. \tag{15}$$

We can rewrite (14) as

$$y_x^\delta = \frac{\mathbb{E}_{Y \sim p_{g,x}^\delta}[Y \exp(-f(Y)/\delta)]}{\mathbb{E}_{Y \sim p_{g,x}^\delta}[\exp(-f(Y)/\delta)]}. \tag{16}$$

This can be interpreted as the expectation in a model in which we first sample $Y$ from the density $p_{g,x}^\delta$, which (assuming that $g$ is minimized at the origin) is centered at $x$ and increasingly peaked for smaller $\delta$, and then *exponentially tilt* by $-f/\delta$, which upweights the samples that lead to smaller values of $f$. Note that we can also interpret (16) from the Bayesian perspective: consider a model whose likelihood is $p_{g,x}^\delta(y) \propto e^{-g((x-y)/\delta)}$, and prior is $\pi(y) \propto e^{-f(y)/\delta}$. In this context, the quantity $y_x^\delta$ represents the *posterior mean*.

In the special case of $g = \|\cdot\|_2^2/(2\lambda)$, observe that $p_{g,x}^\delta$ in (15) is the $N(x, \delta\lambda I)$ density, and (16) leads to

$$y_x^\delta = \frac{\mathbb{E}_{Y \sim N(x,\delta\lambda I)}[Y \exp(-f(Y)/\delta)]}{\mathbb{E}_{Y \sim N(x,\delta\lambda I)}[\exp(-f(Y)/\delta)]}, \tag{17}$$

which recovers the proximal approximation formula in Darbon and Langlois (2021); Osher et al. (2023), who arrived at this result from a different perspective, as explained next.

## 2.2 Viscous Burgers' equation

As discussed previously, the Hopf-Lax formula (4) gives the solution to the Hamilton-Jacobi PDE (3). When $H = \|\cdot\|_2^2/2$, problem (3) reduces to what is known as Burgers' equation, and Laplace's approximation can be understood from the perspective of what is called *viscosity* in the PDE literature. In particular, consider the viscous Burgers' equation (Example 2 in Chapter 4.5.2 of Evans (2010)),

$$\begin{aligned} \partial_t u^\delta + \frac{1}{2}\|\nabla u^\delta\|_2^2 &= \frac{\delta}{2}\Delta u^\delta, \quad t > 0, \\ u^\delta(x,0) &= f(x), \qquad t = 0, \end{aligned} \tag{18}$$

where $\Delta u$ is the Laplacian of $u$ with respect to $x$. By using a change of variables $v^\delta(x,t) = \exp(-u^\delta(x,t)/\delta)$ (also known as the Cole-Hopf transform), problem (18) becomes the heat equation, with the initial condition $v^\delta(x,0) = \exp(-f(x)/\delta)$. We can use the fundamental solution of the heat equation, and translate back to our original parametrization, to yield the solution

$$u^\delta(x,t) = -\delta \log\left( \frac{1}{(2\pi\delta t)^{d/2}} \int \exp\left( \frac{-f(y) - \|x-y\|_2^2/(2t)}{\delta} \right) \mathrm{d}y \right). \tag{19}$$

This approximates the solution $u(x,t)$ in (4), i.e., it approximates the Moreau envelope $f_t$ of $f$ (since, recall, $H = \|\cdot\|_2^2/2$). A seminal result by Crandall and Lions (1984) is that $u^\delta(x,t) \to u(x,t)$ as $\delta \to 0^+$, uniformly over all $x \in \mathbb{R}^d$ and all compact intervals of time $t \geq 0$. Further results are available in Darbon and Langlois (2021); Darbon et al. (2021); Heaton et al. (2023); Osher et al. (2023).

As observed by the aforementioned authors, we can combine (19) with the Moreau gradient formula (9) to obtain an approximation to the proximal map. First we differentiate (19) with respect to $x$,

$$\nabla u^\delta(x,t) = \frac{\int \frac{x-y}{t} \exp\left(\frac{-f(y)-\|x-y\|_2^2/(2t)}{\delta}\right) \mathrm{d}y}{\int \exp\left(\frac{-f(y)-\|x-y\|_2^2/(2t)}{\delta}\right) \mathrm{d}y}.$$

Then based on (9), we approximate $\mathrm{prox}_{tf}(x)$ using $x - t\nabla u^\delta(x,t)$. Observe that

$$x - t\nabla u^\delta(x,t) = \frac{\int y \exp\left(\frac{-f(y)-\|x-y\|_2^2/(2t)}{\delta}\right) \mathrm{d}y}{\int \exp\left(\frac{-f(y)-\|x-y\|_2^2/(2t)}{\delta}\right) \mathrm{d}y}$$

is precisely the self-normalized Laplace approximation in (14) with $g = \|\cdot\|_2^2/(2t)$. Recall, this also has the equivalent form (17), expressed in terms of expectations with respect to $N(x, \delta t I)$.

The fact that Laplace's approximation to the proximal operator can be alternatively derived via viscosity in Burgers' equation is quite interesting (and to emphasize once again, this was the path taken by previous work to arrive at the same formula for the proximal approximation that we get from Laplace's method). This begs the question: for general $g$, is there such a viscosity-like representation for (14)? That is, can we find a viscosity-like modification to the Hamilton-Jacobi PDE (3), with $H^* = g$, whose solution leads to the formula (14)? While viscosity is studied for general Hamilton-Jacobi equations (e.g., Chapter 10 of Evans (2010)), as far as we can tell, the viscous Hamilton-Jacobi equation using a Laplacian diffusion term does not lead to an solution that coincides with Laplace's approximation of the inf convolution in general, in the way that it does when $H = \|\cdot\|_2^2/2$. Investigating whether we can express (14) in terms of a viscosity-like pertubation of the Hamilton-Jacobi equation (3) may be an interesting direction for future investigation.

## 2.3 Smoothed set projection

Returning to the posterior mean formula (17) when $g = \|\cdot\|_2^2/2$, consider taking $f = I_{\mathcal{K}}$, the characteristic function of a set $\mathcal{K} \subseteq \mathbb{R}^d$,

$$I_{\mathcal{K}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{K} \\ \infty & \text{otherwise.} \end{cases}$$

In this case, the quantity being approximated is the proximal operator $y_x = \mathrm{prox}_{I_{\mathcal{K}}}(x)$ of $I_{\mathcal{K}}$ evaluated at $x$, i.e., the projection $y_x = P_{\mathcal{K}}(x)$ of $x$ onto the set $\mathcal{K}$ (assumed unique), concretely

$$P_{\mathcal{K}}(x) = \operatorname*{argmin}_{y \in \mathcal{K}} \|x - y\|_2^2.$$

Introducing the notation $P_{\mathcal{K}}^\delta(x) = y_x^\delta$ for the approximation in (17), note that this simplifies to

$$P_{\mathcal{K}}^\delta(x) = \mathbb{E}_{Y \sim N(x,\delta I)}[Y \mid Y \in \mathcal{K}]. \tag{20}$$

This is highly intuitive: we take an average according to a certain density over $\mathcal{K}$, which ends up being flatter at points far away from $P_{\mathcal{K}}(x)$, and more peaked close to $P_{\mathcal{K}}(x)$. See Figure 3 for an illustration.

If $\mathcal{K}$ is convex then the approximation (20) has the property that $P_{\mathcal{K}}^\delta(x) \in \mathcal{K}$ for any $\delta > 0$; for nonconvex $\mathcal{K}$, this no longer needs to be true, and we could have $P_{\mathcal{K}}^\delta(x)$ lying outside of $\mathcal{K}$. However, convexity is not required in order to guarantee $P_{\mathcal{K}}^\delta(x) \to P_{\mathcal{K}}(x)$ as $\delta \to 0^+$. We only require a mild condition on the boundary of $\mathcal{K}$ in a neighborhood of $P_{\mathcal{K}}(x)$ (see Corollary 2).

## 2.4 Integral convolution

Lastly, we make the simple observation that in the current setting the Laplace approximation brings an inf convolution to an ordinary (integral) convolution. Generalizing from (14), suppose we are interested in

$$h(y_x) = h\Big(\operatorname*{argmin}_y \big\{f(y) + g(x-y)\big\}\Big),$$

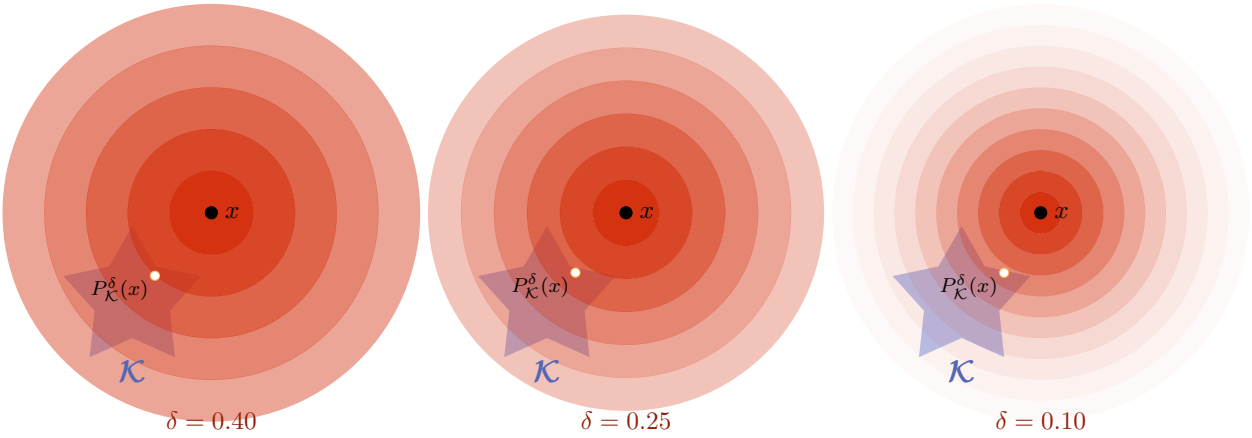



Figure 3: Examples of the Laplace approximation $P_{\mathcal{K}}^{\delta}(x)$ to the projection $P_{\mathcal{K}}(x)$ of a point $x$ onto a set $\mathcal{K}$. In each panel $P_{\mathcal{K}}^{\delta}(x)$ is denoted by a white dot, and is defined by (20) for a particular value of $\delta$. As $\delta$ decreases (from left to right), the conditional density of $Y \,|\, Y \in \mathcal{K}$ becomes increasingly peaked around $P_{\mathcal{K}}(x)$. Each outer red circle about $x$ represents a successive standard deviation $\sqrt{\delta}$ in the contours of the sampling distribution $N(x, \delta I)$.

for a given function $h : \mathbb{R}^d \to \mathbb{R}$ (we also assume that the argmin above is unique). Then the corresponding self-normalized Laplace approximation from (13) is

$$\frac{\int h(y) \exp\left(\frac{-f(y)-g(x-y)}{\delta}\right) \mathrm{d}y}{\int \exp\left(\frac{-f(y)-g(x-y)}{\delta}\right) \mathrm{d}y}. \tag{21}$$

Assuming $g$ is an even function (so that $g(x-y) = g(y-x)$), this can be expressed as

$$\left(\frac{(he^{-f/\delta}) * e^{-g/\delta}}{e^{-f/\delta} * e^{-g/\delta}}\right)(x), \tag{22}$$

where $(u * v)(x) = \int u(y)v(y-x)\,\mathrm{d}y = \int u(y)v(x-y)\,\mathrm{d}y$, the last equality holding for even $v$. As convolution is generally understood as a smoothing operation, this formulation informally lends support to the idea that Laplace method's provides a smooth approximation to $h(y_x)$.

More formally, we can use properties of convolutions to infer about the smoothness of (21) as a function of $x$. If $u, v$ are integrable and $v$ has integrable partial derivatives, then a standard fact (which can be verified using Fourier transforms) is that

$$\frac{\partial(u * v)}{\partial x_i} = u * \frac{\partial v}{\partial x_i}, \quad i = 1, \ldots, n.$$

Applying this to (22) (and using the quotient rule and chain rule as needed), we see that the approximation is differentiable as many times as $g$ is (i.e., it is infinitely differentiable for a choice such as $g = \|\cdot\|_2^2/2$).

# 3 Asymptotic theory

We analyze the asymptotic validity ($\delta \to 0^+$) of the self-normalized version of Laplace's approximation, in a way that generalizes what is known classically (13), which requires twice differentiability (recall Theorem 1). We break our presentation in what follows into two parts, depending on whether the minimizer in question lies in the interior of its domain.

## 3.1 Minimizer in the interior

First, we study the Laplace approximation of a function $\varphi$ whose minimizer is in the interior of its domain. The proof of the next theorem is given in Appendix A.2.

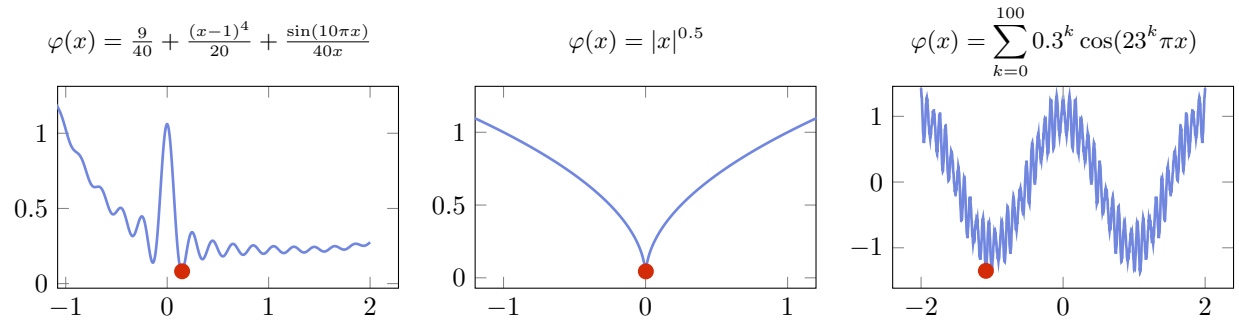



Figure 4: Examples of the Laplace approximation to the minimizer $x^\star$ of $\varphi$. The panels potray different functions $\varphi$, each of which satisfies the local Hölder assumption (23), but is nonconvex and nonsmooth in the traditional sense. In each panel the Laplace approximation is denoted by a red dot, and computed using the left-hand side of (24), with $h$ being the identity map and $\delta = 10^{-6}$.

**Theorem 2.** *Let $\varphi : \mathcal{K} \to \mathbb{R}$ be a continuous function, and assume that $\varphi$ has a unique global minimizer $x^\star$ that lies in the interior of $\mathcal{K}$, admits a bounded sublevel set $\mathcal{S}_\epsilon = \{x \in \mathcal{K} : \varphi(x) \le \varphi(x^\star) + \epsilon\}$ for some $\epsilon > 0$, and satisfies $\int_{\mathcal{K}} \exp(-\varphi(x)/\epsilon)\,\mathrm{d}x < \infty$. Assume that on a neighborhood $U$ of $x^\star$, we have*

$$\varphi(x) - \varphi(x^\star) \le a\|x - x^\star\|_2^q, \quad \text{for all } x \in \mathcal{K} \cap U, \tag{23}$$

*for constants $a, q > 0$. Then for any continuous $h : \mathcal{K} \to \mathbb{R}$ with $\int_{\mathcal{K}} |h(x)| \exp(-d\varphi(x)/(q\epsilon))\,\mathrm{d}x < \infty$,*

$$\frac{\int_{\mathcal{K}} h(x) \exp(-\varphi(x)/\delta)\,\mathrm{d}x}{\int_{\mathcal{K}} \exp(-\varphi(x)/\delta)\,\mathrm{d}x} \;\to\; h(x^\star), \quad \text{as } \delta \to 0^+. \tag{24}$$

The condition (23) is essentially a type of local Hölder continuity assumption on $\varphi$, on a neighborhood $U$ of its minimizer $x^\star$, with an arbitrary exponent $q > 0$.[1] This is considerably weaker than assuming that $\varphi$ is twice differentiable on a neighborhood of $x^\star$, as in Theorem 1. (We note that if $\varphi$ is locally twice continuously differentiable, then it satisfies (23) with $q = 2$, which can be verified using a Taylor expansion. Meanwhile, if $\varphi$ is only locally continuously differentiable, then it satisfies (23) with $q = 1$, as $\|\nabla\varphi\|_2$ has a finite maximum on any compact subset containing its minimizer.)

In order to be able to weaken the assumption from local twice differentiability to local Hölder continuity, the self-normalized aspect of the approximation in (24) is key, because in general the explicit normalizing factors in a statement like (12) would require precise knowledge of the local growth rate of $\varphi(x) - \varphi(x^\star)$. To be clear, in (23), we only need to know that this local growth rate is a power of $\|x - x^\star\|_2$, without needing to know the exponent, in order to compute the approximation (24). Figure 4 gives a few illustrations.

Next, we apply Theorem 2 to an infimal convolution.

**Corollary 1.** *Let $f, g : \mathbb{R}^d \to \mathbb{R}$ be continuous functions, and define $\varphi_x(y) = f(y) + g(x - y)$. Assume that $\varphi_x$ has a unique global minimizer $y_x$, has a bounded sublevel set $\mathcal{S}_\epsilon = \{y \in \mathbb{R}^d : \varphi_x(y) \le \varphi_x(y_x) + \epsilon\}$ for some $\epsilon > 0$, and satisfies $\int \exp(-\varphi_x(y)/\epsilon)\,\mathrm{d}y < \infty$. Assume further that on a neighborhood $U_x$ of $y_x$,*

$$\varphi_x(y) - \varphi_x(y_x) \le a\|y - y_x\|_2^q, \quad \text{for all } y \in U_x, \tag{25}$$

*for constants $a, q > 0$. Then for any continuous $h : \mathbb{R}^d \to \mathbb{R}$ with $\int |h(y)| \exp(-d\varphi_x(y)/(q\epsilon))\,\mathrm{d}y < \infty$,*

$$\frac{\int h(y) \exp\left(\frac{-f(y)-g(x-y)}{\delta}\right) \mathrm{d}y}{\int \exp\left(\frac{-f(y)-g(x-y)}{\delta}\right) \mathrm{d}y} \;\to\; h(y_x), \quad \text{as } \delta \to 0^+.$$

*An important special case is given by taking $h(y) = y_i$, $i = 1, \ldots, d$, for which the above conclusion translates as follows: provided $\int \|y\|_\infty \exp((-f(y) - g(x-y))/\epsilon)\,\mathrm{d}y < \infty$, it holds for $y_x^\delta$ as defined in (14) that*

$$y_x^\delta \to y_x, \quad \text{as } \delta \to 0^+.$$

[1] We say "essentially" because, technically, Hölder continuity is stronger, and would require that $\varphi(x) - \varphi(y) \le a\|x - y\|_2^q$ for all pairs $x, y \in U$, whereas the condition (23) only requires that this inequality to hold when $y = x^\star$.

Once again we note that the local Hölder condition (25) on $\varphi_x$ is quite weak. For example, this condition holds if $f, g$ are each convex on a neighborhood of $y_x$ (since convex functions are locally Lipschitz), or even if $f, g$ are weakly convex on a neighborhood of $y_x$ (see Appendix A.3 for a precise statement and verification of this claim). While (weak) convexity is of central interest in many applications of optimization, we emphasize that no kind of convexity is required for (25). Local Hölder continuity, with a possibly fractional exponent $q$, can be satisfied by fairly exotic and highly nonconvex functions (recall the functions from Figure 4).

Lastly, we note that with a more sophisticated analysis it is likely that the local Hölder condition in (25) could be weakened further into one on the local modulus of continuity, where we require that the modulus of continuity have subexponential growth around zero (this is sufficient to imply that the error term analyzed in step 3 of the proof vanishes; see Appendix A.2). However, we do not pursue such an extension.

## 3.2 Minimizer on the boundary

We consider the case in which the minimizer of $\varphi$ lies on the boundary of its domain. The proof of the next theorem is given in Appendix A.4.

**Theorem 3.** *Under the conditions of Theorem 2, assume instead that the unique global minimizer $x^\star$ of $\varphi$ lies on the boundary of the closed set $\mathcal{K}$. Furthermore, assume that $\mathcal{K}$ is full-dimensional and star-shaped, in a local sense around $x^\star$: precisely, writing $B(x, r) = \{y \in \mathbb{R}^d : \|y - x\|_2 \leq r\}$ for the ball of radius $r$ centered at $x$, we assume that there exists $r_0 > 0$ such that $\mathcal{K} \cap B(x^\star, r_0)$ has positive Lebesgue measure, and*

$$x \in \mathcal{K} \cap B(x^\star, r_0) \implies \alpha x + (1-\alpha)x^\star \in \mathcal{K}, \quad \text{for all } \alpha \in [0, 1]. \tag{26}$$

*Then the same conclusion as in (24) holds.*

Next, we apply Theorem 3 to a projection.

**Corollary 2.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ and $x \in \mathbb{R}^d$ be such that $x$ has a unique projection $y_x = P_{\mathcal{K}}(x)$ onto the closed set $\mathcal{K}$. Assume that there exists $r_0 > 0$ such that $\mathcal{K} \cap B(y_x, r_0)$ has positive Lebesgue measure and condition (26) holds, with $y_x$ in place of $x^\star$. Then for any continuous $h : \mathbb{R}^d \to \mathbb{R}$ with $\int_{\mathcal{K}} |h(y)| \exp(-d\|x - y\|_2^2/2)\, \mathrm{d}y < \infty$,*

$$\frac{\int_{\mathcal{K}} h(y) \exp\left(\frac{-\|x-y\|_2^2}{\delta}\right) \mathrm{d}y}{\int_{\mathcal{K}} \exp\left(\frac{-\|x-y\|_2^2}{\delta}\right) \mathrm{d}y} \to h(y_x), \quad \text{as } \delta \to 0^+.$$

*An important special case is given by taking $h(y) = y_i$, $i = 1, \ldots, d$, for which the above conclusion translates as follows: provided $\int_{\mathcal{K}} \|y\|_\infty \exp(-d\|x - y\|_2^2/2)\, \mathrm{d}y < \infty$, it holds for $P_{\mathcal{K}}^\delta(x)$ as defined in (20) that*

$$P_{\mathcal{K}}^\delta(x) \to P_{\mathcal{K}}(x), \quad \text{as } \delta \to 0^+.$$

The assumptions on the set $\mathcal{K}$ used above, in the theorem and corollary, are not strong. They are met if $\mathcal{K}$ is a convex body (see Appendix A.5 for a precise statement and proof of this claim). Yet the end results will continue to hold well outside of convexity. Fundamentally, the stated assumptions on $\mathcal{K}$ are really used as sufficient conditions to ensure that

$$\mathcal{H}^{d-1}(\mathcal{K} \cap \partial B(x^\star, r)) \geq c r^{d-1}, \quad \text{for all } r \leq r_0, \tag{27}$$

where $c > 0$ is constant, and $\mathcal{H}^{d-1}$ denotes Hausdorff measure of dimension $d - 1$. Informally, this condition says that $\mathcal{K}$ should act in the way we would expect of a "regular" full-dimensional set with nonempty interior, locally around $x^\star$. Indeed, if $\mathcal{K} = \mathbb{R}^d$, then condition (27) is met because $\mathcal{H}^{d-1}(\partial B(x^\star, r)) = c r^{d-1}$ for any $r$ and a constant $c > 0$ depending only on $d$. Generally, one can interpret (27) as requiring the set $\mathcal{K}$ to retain a constant fraction of the surface measure over the boundary of the ball $B(x^\star, r)$, for enough small $r$, which can still be satisfied by a highly nonconvex set $\mathcal{K}$, as long as $\mathcal{K}$ has a somewhat "regular" behavior around its boundary (recall Figure 3). We note that the condition (27) could likely be weakened further with a more sophisticated analysis, but we do not pursue this.

# 4 Monte Carlo sampling

In this section, we briefly discuss how to use sampling to approximate Laplace's approximation (17) of the proximal operator, and more generally (16) for the minimizer of $f + g(x - \cdot)$. We should note at the outset that Monte Carlo sampling is a rich field with many powerful tools, and the ideas we describe here are only very basic. Likely, more advanced tools from the Monte Carlo literature (and even from quasi-Monte Carlo) could be applied to improve accuracy and efficiency.

For the proximal formula (17), we can approximate this using a sample average over Gaussian draws:

$$Y_i \overset{\text{iid}}{\sim} N(x, \delta\lambda I), \; i = 1, \ldots, N, \tag{28}$$

$$y_x^{\delta,N} = \frac{\sum_{i=1}^N Y_i \exp(-f(Y_i)/\delta)}{\sum_{i=1}^N \exp(-f(Y_i)/\delta)}. \tag{29}$$

We note that (29) can also be succinctly written as

$$y_x^{\delta,N} = (Y_1, \ldots, Y_N)^\mathsf{T} \operatorname{softmax}(-f(Y_1)/\delta, \ldots, -f(Y_N)/\delta), \tag{30}$$

where for a vector $v = (v_1, \ldots, v_N) \in \mathbb{R}^N$, we denote $\operatorname{softmax}(v) = (e^{v_1}/\sum_{i=1}^N e^{v_i}, \ldots, e^{v_N}/\sum_{i=1}^N e^{v_i})$. Practical implementations of the softmax operator (such as that in SciPy) commonly shift the exponents in order to avoid overflow, instead computing $\operatorname{softmax}(v) = (e^{v_1-a}/\sum_{i=1}^N e^{v_i-a}, \ldots, e^{v_N-a}/\sum_{i=1}^N e^{v_i-a})$, for a scalar value $a$. Any value will contribute a common factor $e^a$ which cancels in the numerator and denominator; but a careful choice such as $a = \max_{i=1,\ldots,N} v_i$ can lead to be better numerical accuracy (e.g., see Blanchard et al. (2021)). For this reason, it can be advantageous to implement (29) using (30) in practice (to take advantage of built-in shifting for numerical robustness) and the same comment applies to all of the formulae involving exponentially-weighted averages in the remainder of this section.

For general $g$, we can approximate (16) analogously using a sample average over suitable draws:

$$Y_i \overset{\text{iid}}{\sim} p_{g,x}^\delta, \; i = 1, \ldots, N, \tag{31}$$

$$y_x^{\delta,N} = \frac{\sum_{i=1}^N Y_i \exp(-f(Y_i)/\delta)}{\sum_{i=1}^N \exp(-f(Y_i)/\delta)}, \tag{32}$$

where $p_{g,x}^\delta$ is the density in (15). Sampling from $p_{g,x}^\delta$ will really only be possible in certain special cases. For example, aside from the case $g = \|\cdot\|_2^2$, where $p_{g,x}^\delta$ is a Gaussian density, we note that when $g = \|\cdot\|_1$, the density $p_{g,x}^\delta$ is a product of Laplace densities (one for each coordinate). Outside of such special cases, we can use importance sampling, where instead of (31), (32), we compute:

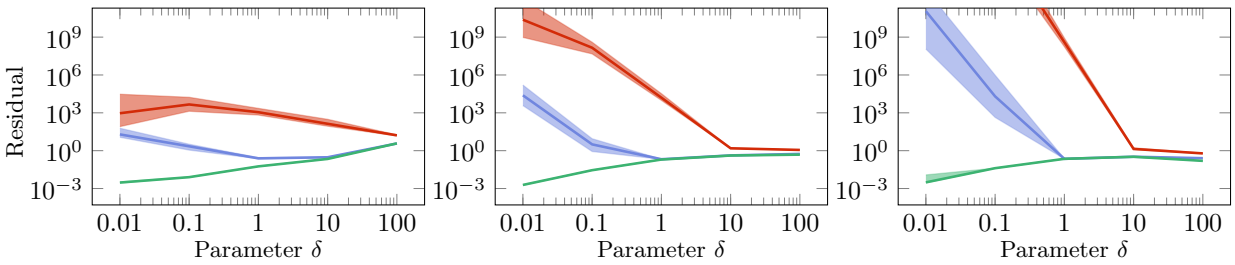
$$Y_i \overset{\text{iid}}{\sim} q, \; i = 1, \ldots, N, \tag{33}$$

$$y_x^{\delta,N} = \frac{\sum_{i=1}^N Y_i w(Y_i) \exp(-f(Y_i)/\delta)}{\sum_{i=1}^N w(Y_i) \exp(-f(Y_i)/\delta)}, \tag{34}$$

where $w(y) = p_{g,x}^\delta(y)/q(y)$, and $q$ is a user-chosen proposal density (whose support contains that of $p_{g,x}^\delta$). In general, a goal in choosing the proposal density $q$ would be to minimize the variance of the weighted sample average in (34). This is a nontrivial task, but various practical solutions have been developed by Hesterberg (1988); Veach (1997); Owen and Zhou (2000), among others.

# 5 Applications and examples

In what follows, we walk through applications of the Laplace approximations proposed and studied above to problems in PDEs and optimization. To be clear, in each case, we do not intend to produce or compete with state-of-the-art solutions for the problem at hand. We only aim to demonstrate the broad applicability and portability of Laplace's approximation, through relatively simple experiments. We focus on low-dimensional problems were sampling is fairly easy (naive Monte Carlo or importance sampling works fairly well). Higher-dimensional problems would call for more advanced sampling techniques.




(a) Dimension $d = 2$, norm $p = 2$ (b) Dimension $d = 2$, norm $p = 5$ (c) Dimension $d = 2$, norm $p = 10$

Figure 5: Residuals for HJ approximation in dimension $d = 2$, for $H = \|\cdot\|_p^p/p$, with $p \in \{2, 5, 10\}$. Each panel shows $N = 10$ samples via red; $N = 10^3$ samples via blue; and $N = 10^5$ samples via green. We can see that as the dimension grows, the errors also grow. When enough samples are used, smaller $\delta$ leads to more accurate solutions. Finally, when not enough samples are used, larger $\delta$ helps to control the errors.

## 5.1 Hamilton-Jacobi equations

We examine the use of Laplace's method for solving the Hamilton-Jacobi (HJ) equation (3). In particular, we study how $\delta$ and the number of samples $N$ used in the approximation (33), (34) influence the quality of the (approximate) HJ solution, for different convex Hamiltonians $H(x) = \|x\|_p^p/p$, $p \in \{1.1, 2, 5, 10\}$, in different dimensions $d \in \{2, 5, 10\}$. We also set $f(x) = \|x\|_1$.

Recall that the exact HJ solution $u(x, t)$ is given in (4). Consider approximating this by first computing $y_x^\delta$ as in (14), and then using $u^\delta(x, t) = f(y_x^\delta) + g(y_x^\delta - x)$, where (to match (4)) we set

$$g(x) = tH^*\left(\frac{x}{t}\right), \quad \text{and} \quad H^*(x) = \frac{1}{q}\|x\|_q^q,$$

for $1/q + 1/p = 1$. For any fixed $t$, we approximate $y_x^\delta$ with $y_x^{\delta,N}$ in (33), (34), where $N$ denotes the number of samples. Then, we calculate

$$u^{\delta,N}(x, t) = f(y_x^{\delta,N}) + g(y_x^{\delta,N} - x), \tag{35}$$

at 1000 uniformly sampled values of $x \in [-10, 10]^d$ and $t \in [10^{-1}, 1]$. The error of the approximation $u^{\delta,N}$ is measured by calculating the magnitude of the HJ residual,

$$r(x, t) = \left|\partial_t u^{\delta,N}(x, t) + H(\nabla u^{\delta,N}(x, t))\right|, \tag{36}$$

and utlimately averaging this over the sampled values of $x$ and $t$.

To compute the Laplace approximation $y_x^{\delta,N}$ in (33), (34) we choose the proposal density $q$ to be uniform over $[-10, 10]^d$. Practically, limiting the domain in this way has little effect (integrating over larger domains lead to similar results). Given $u^{\delta,N}$ in (35), we compute the HJ residual (36) using PyTorch (Paszke et al., 2019)—by tracking the approximation $y_x^{\delta,N}$ within PyTorch's computational graph, the available autograd functionality enables automatic differentiation of $u^{\delta,N}$ with respect to both $x$ and $t$.

Figure 5 shows the median along with the 20th and 80th percentiles of HJ residuals for the Hamiltonians with $p \in \{2, 5, 10\}$ in dimension $d = 2$, over 50 repetitions. As expected, more samples lead to more accurate solutions; when enough samples are used, the error vanishes as $\delta \to 0^+$ (the green curve diminishes at the left end of each plot). Perhaps more interesting is the effect of $\delta$ when *not* enough samples are used: in this case we find that *larger* values of $\delta$ will often lead to more accurate solutions (the blue and red curves *decrease* as $\delta$ increases, for small values of $\delta$). This happens because a larger $\delta$ has a greater regularization effect, via the viscous HJ PDE interpretation discussed in Section 2.2 (and is consistent with what is observed in previous work, such as Chaudhari et al. (2018); Osher et al. (2023)).

Appendix B displays the full set of results over all norms $p$ and dimensions $d$ considered. The behavior is broadly similar to what is observed and described above, but unsurprisingly, higher dimensions lead to larger errors (they would require more samples). Also, $p = 1.1$ acts as somewhat of an exceptional case, as it tends to be more difficult overall. This is probably due to instability in the autodifferentiation calculation used for the residual (36) of $u^{\delta,N}(x, t)$ in (35). Recall, here $g$ is based on the conjugate $H^* = \|\cdot\|_q^q/q$ of $H = \|\cdot\|_p^p/p$, and as $1/p + 1/q = 1$, we have $q \to \infty$ as $p \to 1$.

## 5.2 Proximal point algorithm

We consider the use of Laplace's approximation of the proximal mapping within the context of running the proximal point algorithm on a set of benchmark functions (Hansen et al., 2009), which in all cases except one (the "sphere" function) do not admit analytic proximal maps. We note that the proximal point algorithm (8) with (28), (29) in place of the exact proximal operator, repeats the following update for $k = 1, 2, 3, \ldots$:

$$Y_i \overset{\text{iid}}{\sim} N(x_{k-1}, \delta\lambda I), \; i = 1, \ldots, N, \tag{37}$$

$$x_k = \frac{\sum_{i=1}^N Y_i \exp(-f(Y_i)/\delta)}{\sum_{i=1}^N \exp(-f(Y_i)/\delta)}. \tag{38}$$

This is very intuitive: we explore the space locally by sampling points (37) in a neighborhood of our current iterate $x_{k-1}$, and then we take our next iterate $x_k$ to be a weighted average (38) of these sample, where we exponentially tilt in favor of samples with smaller criterion values.

### 5.2.1 Benchmark functions

In Figure 6, we study the effect of various choices of $\delta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, $N \in \{10, 10^2, 10^3, 10^4\}$ in running the Laplace-proximal point method (37), (38) over functions in the benchmark suite. Throughout we fix $\lambda = 1$. As a reference, in each setting, we plot the lowest criterion value achieved by running gradient descent (GD) over 10,000 iterations, as a dashed black line. To be as favorable possible toward GD, for each setting, we tune over the choice of step size used by GD, as well as a variance level for Gaussian noise to add to the gradient at each iteration, which includes zero noise (usual GD). This Langevin-type modification of GD may help escape local minima in some of the nonconvex benchmark functions. We report the *best* result over all step sizes and noise levels for GD (the one with the lowest criterion value 10,000 iterations) as the dashed black line. All functions in the benchmark are in $d = 10$ dimensions, and obtain a global minimum criterion value of zero at the origin, $x^\star = 0$. We initialize all algorithms at $x_0 = (4, \ldots, 4) \in \mathbb{R}^{10}$, and average all results over 3 repetitions (of random noise generation used in the algorithms).

In the first setting "sphere", the dashed black line is not visible, as GD converges to the global minimum of zero, lying outside of the plotting range. The "sphere" benchmark is the only one in which the criterion is both convex and well-conditioned. Beyond the "sphere" example, we see that Laplace-proximal point (LPP) can compete with and even clearly outperform GD. In two settings, "ellipsoidal" and "discus", LPP starts to outperform GD at a reasonably small number of samples $N$ and reasonably large noise level $\delta$; while in two others, "rosenbrock" and"sharp ridge", it only outperforms GD for larger $N$ and smaller $\delta$. The "weierstrass" example, meanwhile, is different: in contrast to all of the other benchmarks, LPP does not show consistent improvement as $N$ increases and $\delta$ decreases. This is likely due to the fact that this function has a complex landscape with many local minima.

### 5.2.2 Comparison with RGF

We now compare to a well-known zeroth-order method based on Gaussian smoothing, proposed by Nesterov and Spokoiny (2017). This method is based on gradient descent, but approximates the gradient of the criterion $f$ at each iterate $x_{k-1}$ using a finite difference approximation that employs Gaussian perturbations:

$$Y_i \overset{\text{iid}}{\sim} N(x_{k-1}, \delta I), \; i = 1, \ldots, N, \tag{39}$$

$$x_k = x_{k-1} - \frac{\eta}{N} \sum_{i=1}^N \frac{f(Y_i) - f(x_{k-1})}{\delta}, \tag{40}$$

where $\eta > 0$ is a step size. Nesterov and Spokoiny (2017) refer to this algorithm as the *random gradient-free oracle* or RGF. (These authors only consider $N = 1$, but we find that averaging over multiple Gaussian draws tends to improve results).

While both use Gaussian sampling, it is interesting to note that the motivation for RGF is quite different from that for Laplace-proximal point (LPP) in (37), (38). LPP can be interpreted as follows:

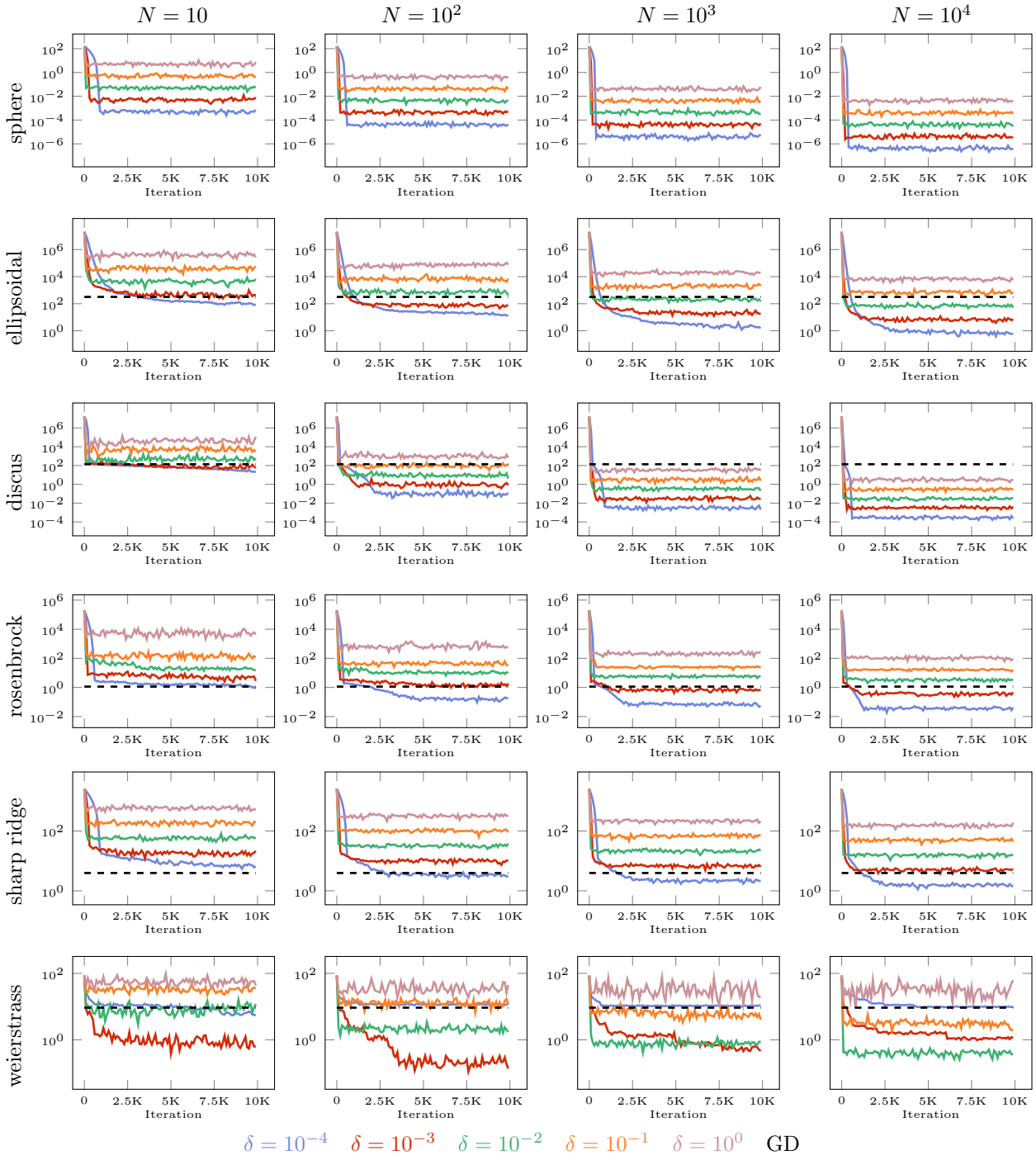



Figure 6: Laplace-proximal point algorithm applied to various benchmark criterion functions. This algorithm repeats the iterations (37), (38) for a particular noise level $\delta$ and number of samples $N$. Each row shows a different benchmark function, each column shows a different number of samples $N \in \{10, 10^2, 10^3, 10^4\}$; and each panel display results for noise levels $\delta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. The result from running gradient descent (GD, tuned over both the choice of step size and added noise level, as explained in the text) is shown as a dashed black line. In the first row, which is a convex and well-conditioned example, GD obtains the global minimum criterion value of zero, and the dashed line is not visible. In all others, the Laplace-proximal point method is able to compete with and even outperform GD.

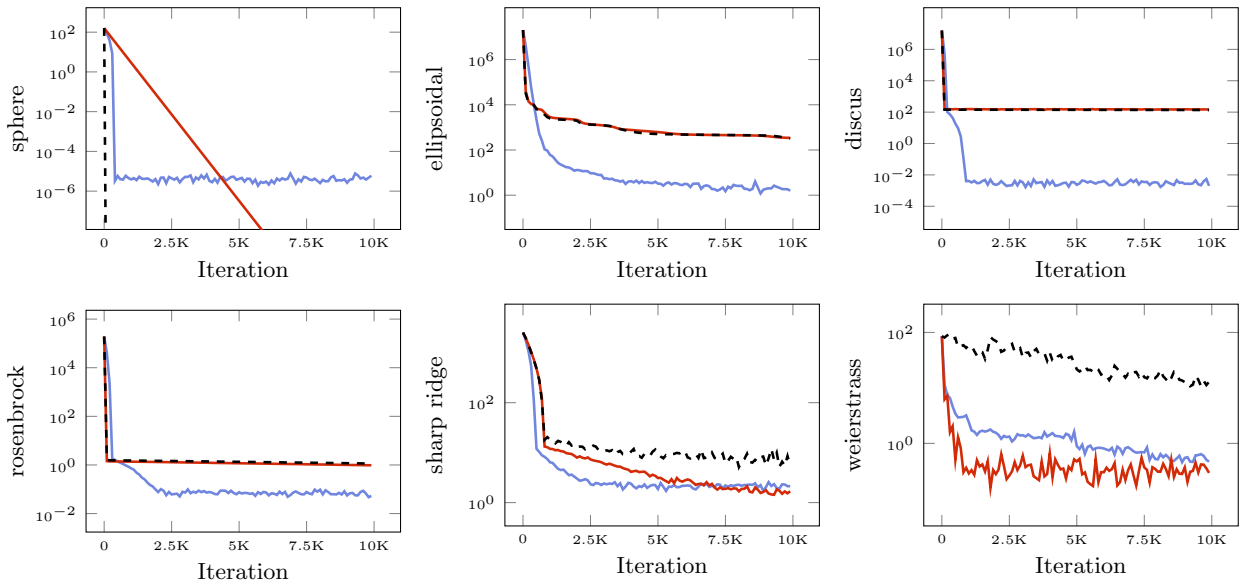



Figure 7: Comparison of random-gradient free oracle (RGF) in red, as in (39), (40); Laplace-proximal point (LPP) in blue, as in (37) (38); and gradient descent (GD) in dotted black; on the same set of benchmark critera from Figure 6. We fix $N = 1000$ samples for RGF and LPP, fix $\lambda = 1$ for LPP, and otherwise allow each algorithm to tune over their respective tuning parameters (noise level $\delta$ for each algorithm, and step size $\eta$ for RGF and GD).

- we first smooth $f$ using its Moreau envelope $f_\lambda$ (recalling that proximal point (8) is nothing more that gradient descent on the Moreau envelope (10));
- we then numerically approximate a gradient step with respect to the Moreau envelope (proximal map) using Laplace's method and Gaussian sampling.

On the other hand, RGF (as Nesterov and Spokoiny (2017) show) can be interpreted as follows:

- we first smooth $f$ using a Gaussian convolution;
- we then numerically approximate a gradient of this convolved function using Gaussian sampling.

Figure 7 compares RGF and LPP on the same set of benchmark criteria as in Figure 6. We fix $N = 1000$ for each algorithm, to equalize their sampling cost. Unlike Figure 6, we now tune LPP over the noise level $\delta$, reporting results for the best noise level (resulting in the smallest criterion value in 10,000 iterations) in each setting. To be as favorable as possible to RGF, we tune it over *both* the noise level $\delta$ and the step size $\eta$, and report results for the best combination of noise level and step size in each benchmark. The results for GD are also shown, where we again tune it over both the noise level and step size, as explained previously. Outside of the convex and well-conditioned "sphere" benchmark, where GD performs best, RGF sometimes improves on GD ("sharp ridge", "weierstrass"), and other times RGF and GD perform quite similarly ("ellipsoidal", "discus", "rosenbrock"). LPP is competitive overall, sometimes improving on RGF ("ellipsoidal", "discus", "rosenbrock") and sometimes not ("sphere", "weierstrass").

## 5.3 Bregman proximal gradient descent

Lastly, we consider the use of Laplace's method to approximate a Bregman proximal map within the context of Bregman proximal gradient descent (BPGD). In particular, we examine a variation on BPGD proposed by Bauschke et al. (2017), which uses an elegant majorization scheme to handle composite criterion functions where the differentiable part lacks a Lipschitz continuous gradient. We focus on a regularized Poisson linear inverse problem, as studied in their paper, where we aim to minimize

$$f(x) = D_\phi(b, Ax) + \mu\|x\|_1, \tag{41}$$

over $\mathbb{R}^d_+$, where $b \in \mathbb{R}^n_+$, $A \in \mathbb{R}^{n\times d}_{++}$, $\mu \geq 0$ is a regularization parameter, $\phi$ is the Boltzmann-Shannon entropy $\phi(x) = \sum_{i=1}^d x_i \log x_i$, and $D_\phi$ denotes its corresponding Bregman divergence,

$$\begin{aligned} D_\phi(b, Ax) &= \phi(b) - \phi(Ax) - \nabla\phi(Ax)^\mathsf{T}(b - Ax) \\ &= \underbrace{\sum_{i=1}^n \left( b_i \log \frac{b_i}{(Ax)_i} + b_i - (Ax)_i \right)}_{d(x)}. \end{aligned} \tag{42}$$

To minimize (41), we consider the BPGD algorithm of Bauschke et al. (2017) and choose as the majorizer the Burg entropy $h(x) = -\sum_{i=1}^d \log x_i$. For the Poisson inverse problem, the Burg entropy is a specially-crafted majorizer which satisfies two key properties:

- it majorizes the Bregman divergence in (42), in the sense that $Lh - d$ is convex for some $L > 0$, which Bauschke et al. (2017) prove is true for any choice $L \geq \|b\|_1$;
- it admits a closed-form Bregman proximal update (in the second line below),

$$x_k = \operatorname*{argmin}_x \left\{ \mu\|x\|_1 + \nabla d(x_{k-1})^\mathsf{T} x + \frac{1}{\eta} D_h(x, x_{k-1}) \right\} \tag{43}$$

$$= \left[ \frac{x_{k-1,i}}{1 + \eta(\mu + \nabla_i d(x_{k-1})) \cdot x_{k-1,i}} \right]_{i=1}^d, \tag{44}$$

  were $\eta \in (0, 1/L)$ is a step size.

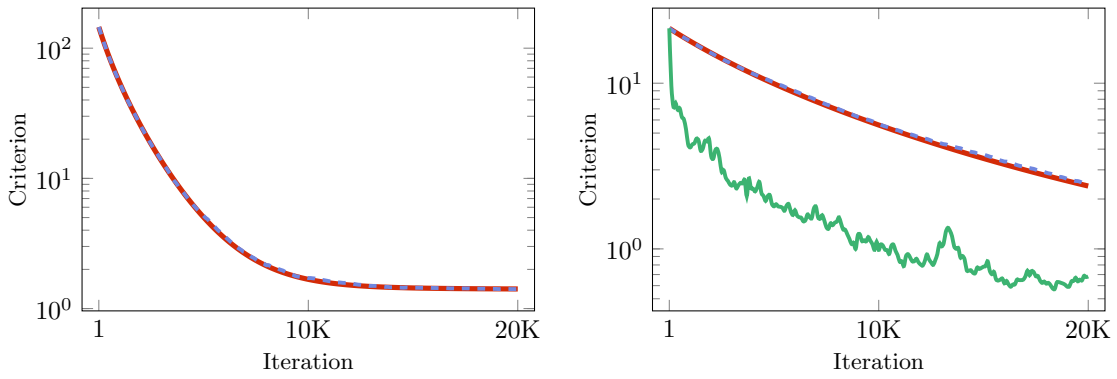
The left panel of Figure 8 shows an example of running the exact BPGD update (44), and Laplace's method to approximate the general form in (43), which does not use knowledge of the fact that for the Burg entropy $h$ the Bregman proximal mapping is exact. For the latter, because the Bregman proximal mapping separates into a minimization problem per coordinate (due to the separability of the Burg entropy itself), we can then approximate each of these univariate infimums using Laplace's method, and run importance sampling as in (33), (34) but on each coordinate separately, with $q$ being the uniform density over $[10^{-6}, 50]$.

In the left panel, we set $n = d = 5$, and generate $A \in \mathbb{R}^{5\times 5}_{++}$ by sampling its entries independently from a uniform distribution on $[1, 2]$. We generate $\bar{x} \in \mathbb{R}^5$ by sampling its entries independently from a uniform on $[5, 6]$, and randomly set half of these to 0. Then, we generate $b \in \mathbb{R}^5_+$ by sampling its entries independently from Poisson distributions with means $(A\bar{x})_i$, $i = 1, \ldots, 5$. We fix $\mu = 10^{-3}$ for the regularization parameter, $\eta = 10^{-5}$ for the step size, $\delta = 2 \cdot 10^{-3}$ for the level of noise, and $N = 5 \cdot 10^4$ for the number of samples. The figure shows the convergence curves (criterion values per iteration) for the exact and Laplace-based BPGD methods. These look identical, as should be expected for such a large number of samples $N$ and small noise level $\delta$ in $d = 5$ dimensions.

Meanwhile, the right panel of Figure 8 shows a different example setting in which $A$ is ill-conditioned and BPGD with the Burg entropy (whether exact or Laplace-based) converges slowly as a result. The setup is the same except that we set $A = aa^\mathsf{T}$, where $a \in \mathbb{R}^5_{++}$ has entries sampled independently from a uniform on $[0, 1]$. Now, in addition to the exact BPGD and Laplace-BPGD methods that use the Burg entropy, we consider a third method: we use the *variable-metric* majorizer $h(x) = -\sum_{i=1}^d \log(Ax)_i$. This does not lead to an exact proximal update, but nonetheless (43) can be approximated using Laplace's method and importance sampling, as in (33), (34), where we take $q$ to be the exponential distribution, arising from the term $\exp(-\mu\|x\|_1)$ that appears in the integral. As we can see in the right panel of the figure, such a variable-metric Laplace-BPGD approach converges more rapidly than standard BPGD, because the variable-metric approach has effectively transformed the parameter space to account for the underlying geometry.

# 6 Discussion

In this work, we study connections between Laplace's approximation and infimal convolutions, with an eye toward solving optimization problems (using proximal-type methods) as well as Hamilton-Jacobi PDEs. Our theory reflects the broad applicability of Laplace's method for approximating infimal convolutions, allowing




(a) Well-conditioned (b) Ill-conditioned

Figure 8: Comparison of BPGD approaches in Poisson linear inverse problems where the underlying linear transform is well-conditioned or ill-conditioned. Here we plot exact BPGD with Burg entropy in red, as in (44); Laplace-BPGD with Burg entropy in dotted blue, as in (43) where $h(x) = -\sum_{i=1}^{d} \log x_i$; and Laplace-BPGD with the variable-metric majorizer in green, as in (43) where $h(x) = -\sum_{i=1}^{d} \log(Ax)_i$. The Laplace methods use $\delta = 2 \cdot 10^{-3}$ as the noise level and $N = 5 \cdot 10^4$ as the number of samples. We can see that the Laplace-BPGD method with Burg majorizer tracks its exact counterpart closely. In the ill-conditioned setting, the variable-metric majorizer accelerates convergence.

us to handle nonconvex functions, and settings where a minimizer lies on the boundary of the domain (such as projections). Our experiments demonstrate the versatility of the Laplace approximation in a few different problem areas, which span optimization and PDEs. Practically, the challenge of sampling in high-dimensional spaces remains a significant hurdle, and one that we do not attempt to address at all. Any implementation based on Laplace's method which strives for both precision and efficiency should likely invest in more advanced sampling techniques. Future work may be able to provide samplers tailored to particular environments and end-goals in optimization and PDEs. An open-source repository with code to reproduce our experiments is available at `https://github.com/mines-opt-ml/laplace-inf-conv`.


## Acknowledgments

We thank Lawrence Craig Evans and Luis Tenorio for helpful discussions. This work was supported in part by Office of Naval Research MURI grant N00014-20-1-2787, to RJT and SO; Air Force Office of Scientific Research MURI grant FA9550-18-502, to SO; and National Science Foundation grant DMS211074, to SWF.


# References


Francis Bach. Approximating integrals with Laplace's method. Machine Learning Research Blog, `https://francisbach.com/laplace-method/`, 2021.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017.

Richard Bellman and William Karush. On a new functional transform in analysis: the maximum transform. *Bulletin of the American Mathematical Society*, 67(5):501–53, 1961.

Richard Bellman and William Karush. Mathematical programming and the maximum transform. *Journal of the Society for Industrial and Applied Mathematics*, 10(3):550–567, 1962a.

Richard Bellman and William Karush. On the maximum transform and semigroups of transformations. *Bulletin of the American Mathematical Society*, 68(5):516–518, 1962b.

Richard Bellman and William Karush. On the maximum transform. *Journal of Mathematical Analysis and Applications*, 6(1):67–74, 1963.

Pierre Blanchard, Desmond J. Higham, and Nicholas J. Higham. Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis*, 41(4):2311–2330, 2021.

Steve Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternative direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillaume Carlier. Deep relaxation: Partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5:1–30, 2018.

Patrick L. Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

Michael G. Crandall and Pierre-Louis Lions. Viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42, 1983.

Michael G. Crandall and Pierre-Louis Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Mathematics of Computation*, 43(167):1–19, 1984.

Michael G. Crandall, Lawrence C. Evans, and Pierre-Louis Lions. Some properties of viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 282(2):487–502, 1984.

Jérôme Darbon and Gabriel P. Langlois. On Bayesian posterior mean estimators in imaging sciences and Hamilton-Jacobi partial differential equations. *Journal of Mathematical Imaging and Vision*, 63(7):821–854, 2021.

Jérôme Darbon, Gabriel P. Langlois, and Tingwei Meng. Connecting Hamilton-Jacobi partial differential equations with maximum a posteriori and posterior mean estimators for some non-convex priors. *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision*, pages 1–25, 2021.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauerd, and Philipp Hennig. Laplace redux — effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems*, 2021.

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the Sherrington-Kirkpatrick Gibbs measure via algorithmic stochastic localization. In *Annual Symposium of Foundations of Computer Science*, 2022.

Lawrence C. Evans. On solving certain nonlinear partial differential equations by accretive operator methods. *Israel Journal of Mathematics*, 36(3):225–247, 1980.

Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society: Graduate Studies in Mathematics, second edition, 2010.

Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, revised edition, 2015.

Werner Fenchel. Convex cones, sets and functions. Lecture Notes, Princeton University, 1951.

Nikolaus Hansen, Steffen Finck, Raymond Ros, and Anne Auger. Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. HAL Id: inria-00362633, 2009.

Howard Heaton, Samy Wu Fung, and Stanley Osher. Global solutions to nonconvex problems by evolution of Hamilton-Jacobi PDEs. *Communications on Applied Mathematics and Computation*, pages 1–21, 2023.

Tim Hesterberg. *Advances in Importance Sampling*. PhD thesis, Department of Statistics, Stanford University, 1988.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Robert E. Kass, Luke Tierney, and Joseph B. Kadane. Laplace's method in Bayesian analysis. In *Statistical Multiple Integration*, pages 89–100. AMS eBooks: Contemporary Mathematics, 1991.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598): 671–680, 1983.

Pierre-Simon Laplace. Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie royale des sciences de Paris*, VI(621):27–65, 1774.

Cécile Louchet. *Modèles variationnels et bayésiens pour le débruitage d'images: de la variation totale vers les moyennes non-locales*. PhD thesis, Université René Descartes-Paris V, 2008.

Cécile Louchet and Lionel Moisan. Posterior expectation of the total variation model: Properties and experiments. *SIAM Journal on Imaging Sciences*, 6(4):2640–2684, 2013.

J. Matyas. Random optimization. *Automation and Remote Control*, 26(2):246–25, 1965.

Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.

Andrea Montanari. Sampling, diffusions, and stochastic localization. arXiv: 2305.10690, 2023.

Jean Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes rendus des séances de l'Académie des Sciences de Paris*, 255:2897–2899, 1962.

Jean Jacques Moreau. Inf-convolution des fonctions numériques sur un espace vectoriel. *Comptes rendus des séances de l'Académie des Sciences de Paris*, 256:5047–5049, 1963.

Jean Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

Jean Jacques Moreau. Inf-convolution, sous-additivité, convexité des fonctions numerique. *Journal de Mathématiques Pures et Appliquées*, 49:109–154, 1970.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Stanley Osher, Howard Heaton, and Samy Wu Fung. A Hamilton-Jacobi-based proximal operator. *Proceedings of the National Academy of Sciences*, 120(14):e2220469120, 2023.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

Neil Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Machine Learning*, 1(3): 123–231, 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

R. Tyrrell Rockafellar. *Convex Functions and Dual Extremum Problems*. PhD thesis, Department of Mathematics, Harvard University, 1963.

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

R. Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*. Springer, 2009. Third printing.

Ernest K. Ryu and Wotao Yin. *Large-Scale Convex Optimization via Monotone Operators*. Cambridge University Press, 2022.

Zhenming Shun and Peter McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B*, 57(4):749–760, 1995.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, and Stefano Ermon. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations*, 2021.

Thomas Strömberg. *A Study of the Operation of Infimal Convolution*. PhD thesis, Department of Mathematics, Luleå University of Technology, 1994.

Tim Veach. *Advances in Importance Sampling*. PhD thesis, Robust Monte Carlo Methods for Light Transport Simulation, 1997.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, 2011.

# A Proofs

## A.1 Proof of Theorem 1

For ease of reference, we state the theorem before its proof.

**Theorem 1.** *Let $\varphi : \mathcal{K} \to \mathbb{R}$ be continuous over a compact set $\mathcal{K} \subseteq \mathbb{R}^d$. Assume that $\varphi$ has a unique global minimizer $x^\star$ in the interior of $\mathcal{K}$, and $\varphi$ is twice continuously differentiable on a neighborhood of $x^\star$, with strictly positive definite Hessian $\nabla^2\varphi(x^\star)$. Then for any continuous function $h : \mathcal{K} \to \mathbb{R}$,*

$$\sqrt{\det(\nabla^2\varphi(x^\star))}\ \frac{\exp(\varphi(x^\star)/\delta)}{(2\pi\delta)^{d/2}} \int_{\mathcal{K}} h(x)\exp(-\varphi(x)/\delta)\,\mathrm{d}x \ \to\ h(x^\star), \quad \text{as } \delta \to 0^+. \tag{45}$$

*This conclusion extends to the case where $\mathcal{K}$ is not compact (e.g., $\mathcal{K} = \mathbb{R}^d$) provided there exists $\epsilon > 0$ such that the sublevel set $\mathcal{S}_\epsilon = \{x \in \mathcal{K} : \varphi(x) \le \varphi(x^\star) + \epsilon\}$ is bounded and $\int_{\mathcal{K}} |h(x)| \exp(-d\varphi(x)/(2\epsilon))\,\mathrm{d}x < \infty$.*

*Proof.* Without loss of generality, assume that $x^\star = 0$, $\varphi(x^\star) = 0$, and the Hessian of $\varphi$ at $x^\star$ is the identity matrix, i.e., $\nabla^2\varphi(x^\star) = I$. The first equality $x^\star = 0$ can be achieved by shifting the domain, and the second $\varphi(x^\star) = 0$ can be achieved by centering the function, which contributes the factor of $\exp(\varphi(x^\star/\delta))$ to the left-hand side in (45). The Hessian condition can be obtained via a change of variables, and this contributes the factor $\sqrt{\det(\nabla^2\varphi(x^\star))}$ to the left-hand side in (45). Now define

$$I_\delta = \frac{1}{(2\pi\delta)^{d/2}} \int_{\mathcal{K}} h(x)\exp(-\varphi(x)/\delta)\,\mathrm{d}x, \quad \text{for } \delta > 0.$$

We seek to show that

$$\lim_{\delta \to 0^+} I_\delta = h(0).$$

We proceed by separately considering whether the set $\mathcal{K}$ is compact or not.

**Compact case.** Suppose $\mathcal{K}$ is compact. Following Bach (2021), define the function $g\colon \mathcal{K} \to \mathbb{R}$ by

$$g(x) = \begin{cases} \dfrac{\varphi(x)}{\|x\|_2^2} & \text{if } x \neq 0, \\ 1/2 & \text{otherwise.} \end{cases}$$

First note that $g$ is continuous at each $x \neq 0$ since $\varphi$ and $\|\cdot\|_2$ are both continuous (and the denominator is nonzero). Moreover, we can show that $g$ is continuous at $x = 0$ via a second-order Taylor expansion of $\varphi$:

$$\varphi(x) = \varphi(x^\star) + \nabla\varphi(x^\star)^\mathsf{T} x + \frac{1}{2}x^\mathsf{T}\nabla^2\varphi(x^\star)x + R(x) = \frac{1}{2}\|x\|_2^2 + R(x),$$

where the remainder term is such that $R(x)/\|x\|_2^2 \to 0$ as $x \to 0$, and therefore,

$$\lim_{x\to 0} g(x) = \lim_{x \to 0} \frac{\frac{1}{2}\|x\|_2^2 + R(x)}{\|x\|_2^2} = \frac{1}{2} + 0 = g(0).$$

Hence we have shown $g$ is continuous on all of $\mathcal{K}$.

Using the definition of $g$, the integral $I_\delta$ may be equivalently written as

$$\begin{aligned} I_\delta &= \frac{1}{(2\pi\delta)^{d/2}} \int_{\mathcal{K}} h(x) \cdot \exp\Big(-\|x\|_2^2 \cdot g(x)/\delta\Big)\,\mathrm{d}x \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \mathbb{1}_{\mathcal{K}}(\sqrt{\delta}y) \cdot h(\sqrt{\delta}y) \cdot \exp\Big(-\|y\|_2^2 \cdot g(\sqrt{\delta}y)\Big)\,\mathrm{d}y, \end{aligned} \tag{46}$$

where the second line follows from the change of variables $x = \sqrt{\delta}y$ and $\mathbb{1}_{\mathcal{K}}$ is the indicator function of $\mathcal{K}$,

$$\mathbb{1}_{\mathcal{K}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{K} \\ 0 & \text{otherwise.} \end{cases}$$

For each $y$, the absolute value of the integrand in (46) is upper bounded by

$$\left(\max_{x\in\mathcal{K}} |h(x)|\right) \cdot \exp\left( - \|y\|_2^2 \cdot \left(\min_{x\in\mathcal{K}} g(x)\right)\right). \tag{47}$$

By compactness of $\mathcal{K}$ and continuity of $h$, the max term is finite. By compactness of $\mathcal{K}$, continuity of $g$, and the fact that $g$ is strictly positive on $\mathcal{K}$, the min term is strictly positive. These facts imply

$$\int_{\mathbb{R}^d} \left(\max_{x\in\mathcal{K}} |h(x)|\right) \cdot \exp\left( - \|y\|_2^2 \cdot \left(\min_{x\in\mathcal{K}} g(x)\right)\right) \mathrm{d}y < \infty,$$

so the integrand in (46) is dominated by an integrable function (47). By the dominated convergence theorem,

$$\begin{aligned}
\lim_{\delta\to 0^+} I_\delta &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \lim_{\delta\to 0^+} \mathbb{1}_{\mathcal{K}}(\sqrt{\delta}y) \cdot h(\sqrt{\delta}y) \cdot \exp\left( - \|y\|_2^2 \cdot g(\sqrt{\delta}y)\right) \mathrm{d}y \\
&= h(0) \cdot \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \exp(-\|y\|_2^2/2)\, \mathrm{d}y \\
&= h(0),
\end{aligned}$$

completing the proof of the compact case.

**Noncompact case.** Suppose $\mathcal{K}$ is not compact. In this case, we decompose the integral $I_\delta$ into two parts:

$$I_\delta = \underbrace{\frac{1}{(2\pi\delta)^{d/2}} \int_{\mathcal{S}_\epsilon} h(x) \exp(-\varphi(x)/\delta)\, \mathrm{d}x}_{I_{1,\delta}} + \underbrace{\frac{1}{(2\pi\delta)^{d/2}} \int_{\mathcal{K}\setminus\mathcal{S}_\epsilon} h(x) \exp(-\varphi(x)/\delta)\, \mathrm{d}x}_{I_{2,\delta}}.$$

Since $\varphi$ is continuous, the set $\mathcal{S}_\epsilon$ is closed. Thus, because it is also bounded (by assumption), the set $\mathcal{S}_\epsilon$ is compact and the arguments from the last case show $I_{1,\delta} \to h(0)$ as $\delta \to 0$. To complete the proof, it remains to show $I_{2,\delta} \to 0$ as $\delta \to 0^+$. Bringing the leading term inside the integral and taking absolute values reveals:

$$|I_{2,\delta}| \le \int_{\mathbb{R}^d} \underbrace{\mathbb{1}_{\mathcal{K}\setminus\mathcal{S}_\epsilon}(x) \cdot \frac{|h(x)|}{(2\pi\delta)^{d/2}} \cdot \exp(-\varphi(x)/\delta)}_{f_\delta(x)}\, \mathrm{d}x.$$

Furthermore, for $x \in \mathcal{K} \setminus \mathcal{S}_\epsilon$,

$$\begin{aligned}
\frac{\mathrm{d}f_\delta(x)}{\mathrm{d}\delta} &= \frac{|h(x)|}{(2\pi\delta)^{d/2}} \cdot \left(\frac{\varphi(x)}{\delta^2} - \frac{d}{2\delta}\right) \cdot \exp(-\varphi(x)/\delta) \\
&\ge \frac{|h(x)|}{(2\pi\delta)^{d/2}} \cdot \left(\frac{\epsilon}{\delta^2} - \frac{d}{2\delta}\right) \cdot \exp(-\varphi(x)/\delta).
\end{aligned}$$

If $\delta \le 2\epsilon/d$, then the lower bound in the last line above is nonnegative; in other words, we have shown that for $\delta \le 2\epsilon/d$, the quantity $f_\delta(x)$ is nonincreasing as $\delta$ decreases, at each $x \in \mathcal{K} \setminus \mathcal{S}_\epsilon$. As $f_\delta \ge 0$, we conclude that $f_\delta$ is dominated by $f_{2\epsilon/d}$, which is integrable by assumption, and another application of the dominated convergence theorem therefore yields

$$\begin{aligned}
\lim_{\delta\to 0^+} |I_{2,\delta}| &\le \int_{\mathbb{R}^d} \lim_{\delta\to 0^+} f_\delta(x)\, \mathrm{d}x \\
&\le \int_{\mathcal{K}\setminus\mathcal{S}_\epsilon} \lim_{\delta\to 0^+} \frac{|h(x)|}{(2\pi\delta)^{d/2}} \cdot \exp(-\epsilon/\delta)\, \mathrm{d}x \\
&= 0,
\end{aligned}$$

where the last line holds since $\delta^{-d/2} \exp(-\epsilon/\delta) \to 0$ as $\delta \to 0^+$. This completes the proof of the noncompact case, and the theorem. □

## A.2 Proof of Theorem 2

For ease of reference, we state the theorem before its proof.

**Theorem 2.** *Let $\varphi : \mathcal{K} \to \mathbb{R}$ be a continuous function, and assume that $\varphi$ has a unique global minimizer $x^\star$ that lies in the interior of $\mathcal{K}$, admits a bounded sublevel set $\mathcal{S}_\epsilon = \{x \in \mathcal{K} : \varphi(x) \le \varphi(x^\star) + \epsilon\}$ for some $\epsilon > 0$, and satisfies $\int_{\mathcal{K}} \exp(-\varphi(x)/\epsilon)\,\mathrm{d}x < \infty$. Assume that on a neighborhood $U$ of $x^\star$, we have*

$$\varphi(x) - \varphi(x^\star) \le a\|x - x^\star\|_2^q, \quad \text{for all } x \in \mathcal{K} \cap U, \tag{48}$$

*for constants $a, q > 0$. Then for any continuous $h : \mathcal{K} \to \mathbb{R}$ with $\int_{\mathcal{K}} |h(x)| \exp(-d\varphi(x)/(q\epsilon))\,\mathrm{d}x < \infty$,*

$$\frac{\int_{\mathcal{K}} h(x) \exp(-\varphi(x)/\delta)\,\mathrm{d}x}{\int_{\mathcal{K}} \exp(-\varphi(x)/\delta)\,\mathrm{d}x} \;\to\; h(x^\star), \quad \text{as } \delta \to 0^+. \tag{49}$$

*Proof.* As before, we assume without loss of generality that $x^\star = 0$ and $\varphi(x^\star) = 0$. We also assume that the local Hölder bound in (48) holds on all of $\mathcal{S}_\epsilon$; this is possible because $\{\mathcal{S}_t : t \le \epsilon\}$ is a family of compact sets which are decreasing according to the partial ordering given by set inclusion, and $\mathcal{S}_0 = \{x^\star\}$, thus if needed we can just redefine $\epsilon$ to be small enough such that $\mathcal{S}_\epsilon \subseteq \mathcal{K} \cap U$. Now let

$$p^\delta(x) = \frac{\exp(-\varphi(x)/\delta)}{\int_{\mathcal{K}} \exp(-\varphi(x)/\delta)\,\mathrm{d}x}.$$

For $\delta \le \epsilon$, note that the denominator here is finite because the integrand is upper bounded by $\exp(-\varphi(x)/\epsilon)$, which is integrable by assumption. Furthermore, the denominator is positive because the integrand is lower bounded by $\exp(-a\|x - x^\star\|_2^q) > 0$ on $\mathcal{S}_\epsilon$. Hence we have shown $0 < \int_{\mathcal{K}} \exp(-\varphi(x)/\delta)\,\mathrm{d}x < \infty$, which means that $p^\delta$ is a well-defined probability measure on $\mathcal{K}$. We now seek to prove

$$\lim_{\delta \to 0^+} \int_{\mathcal{K}} h(x) p^\delta(x)\,\mathrm{d}x = h(0),$$

which holds if

$$\lim_{\delta \to 0^+} \underbrace{\int_{\mathcal{K}} |h(x) - h(0)| p^\delta(x)\,\mathrm{d}x}_{J_\delta} = 0.$$

We split the integral $J_\delta$ into two parts:

$$J_\delta = \underbrace{\int_{\mathcal{K} \cap B(0,\tau)} |h(x) - h(0)| p^\delta(x)\,\mathrm{d}x}_{J_{1,\delta,\tau}} + \underbrace{\int_{\mathcal{K} \setminus B(0,\tau)} |h(x) - h(0)| p^\delta(x)\,\mathrm{d}x}_{J_{2,\delta,\tau}},$$

where $B(0,\tau) = \{x \in \mathbb{R}^d : \|x\|_2 \le \tau\}$ is the ball of radius $\tau > 0$ centered at the origin, and the radius $\tau$ is yet to be specified. Let $\mu > 0$ be given. It suffices to verify the existence of $\tau > 0$ and $\delta_0 > 0$ such that

$$J_{1,\delta,\tau} \le \frac{\mu}{2} \quad \text{and} \quad J_{2,\delta,\tau} \le \frac{\mu}{2}, \quad \text{for } \delta \le \delta_0.$$

This would lead to $J_\delta = J_{1,\delta,\tau} + J_{2,\delta,\tau} \le \mu$ for all $\delta \le \delta_0$, implying the desired convergence $J_\delta \to 0$ as $\delta \to 0^+$. The rest of this proof is structured as follows. We bound $J_{1,\delta,\tau} \le \mu/2$ by choosing $\tau$ to be sufficiently small (step 1). This bound holds for any $\delta > 0$. For this same $\tau$, we upper bound $J_{2,\delta,\tau}$ by a quantity that depends on $\delta$ (step 2). We then show that the numerator of this upper bound converges to zero (step 3), whereas the denominator is bounded below by a positive constant as $\delta \to 0^+$ (step 4). Together, steps 3 and 4 imply the existence of $\delta_0$ such that $J_{2,\delta,\tau} \le \mu/2$ for all $\delta \le \delta_0$, which will complete the proof.

**Step 1: bounding the integral $J_{1,\delta}$.** As $h$ is continuous, we can choose $\tau$ so that $|h(x) - h(0)| \le \mu/2$ for all $x \in B(0,\tau)$, which makes

$$J_{1,\delta,\tau} \le \frac{\mu}{2} \cdot \int_{\mathcal{K} \cap B(0,\tau)} p^\delta(x)\,\mathrm{d}x \le \frac{\mu}{2}.$$

**Step 2: bounding the integral $J_{2,\delta}$.** Observe

$$J_{2,\delta,\tau} = \frac{\delta^{-d/q} \cdot \int_{\mathcal{K}\setminus B(0,\tau)} |h(x)-h(0)| \exp(-\varphi(x)/\delta)\,\mathrm{d}x}{\delta^{-d/q} \cdot \int_{\mathcal{K}} \exp(-\varphi(x)/\delta)\,\mathrm{d}x} \leq \frac{\delta^{-d/q} \cdot \int_{\mathcal{K}\setminus B(0,\tau)} |h(x)-h(0)| \exp(-\varphi(x)/\delta)\,\mathrm{d}x}{\delta^{-d/q} \cdot \int_{\mathcal{K}\cap B(0,\tau)} \exp(-\varphi(x)/\delta)\,\mathrm{d}x}. \tag{50}$$

The motivation for introducing the factor of $\delta^{-d/q}$ in the numerator and denominator above is a change of variables that will be used in step 4.

**Step 3: analyzing the numerator in (50).** Let $\epsilon_0 > 0$ be such that $\mathcal{S}_{\epsilon_0} \subseteq \mathcal{K} \cap B(0,\tau)$. Such a value of $\epsilon_0$ always exists because, similar to an argument given earlier, $\{\mathcal{S}_t : t \leq \epsilon\}$ is a family of nested compact sets with $\mathcal{S}_0 = \{x^\star\}$. We assume without loss of generality that $\epsilon_0 \leq \epsilon$ (otherwise we just make $\epsilon_0$ smaller). Now rewrite the numerator in (50) as

$$\delta^{-d/q} \cdot \int_{\mathcal{K}\setminus B(0,\tau)} |h(x)-h(0)| \exp(-\varphi(x)/\delta)\,\mathrm{d}x = \int_{\mathbb{R}^d} \underbrace{\mathbb{1}_{\mathcal{K}\setminus B(0,\tau)}(x) \cdot \frac{|h(x)-h(0)|}{\delta^{d/q}} \cdot \exp(-\varphi(x)/\delta)}_{f_\delta(x)}\,\mathrm{d}x.$$

Similar to a calculation given at the end of the proof of Theorem 1 in Appendix A.1, for $x \in \mathcal{K} \setminus B(0,\tau)$,

$$\frac{\mathrm{d}f_\delta(x)}{\mathrm{d}\delta} = \frac{|h(x)-h(0)|}{\delta^{d/q}} \cdot \left(\frac{\varphi(x)}{\delta^2} - \frac{d}{q\delta}\right) \cdot \exp(-\varphi(x)/\delta) \geq \frac{|h(x)-h(0)|}{\delta^{d/q}} \cdot \left(\frac{\epsilon_0}{\delta^2} - \frac{d}{q\delta}\right) \cdot \exp(-\varphi(x)/\delta),$$

where the last line holds since $\varphi(x) \geq \epsilon_0$ (which is implied by our choice of $\epsilon_0$ such that $\mathcal{S}_{\epsilon_0} \subseteq B(0,\tau)$). Thus for all $\delta \leq q\epsilon_0/d$, the quantity $f_\delta(x)$ is nonincreasing as $\delta$ decreases, at each $x \in \mathcal{K} \setminus B(0,\tau)$. As $f_\delta \geq 0$, and $f_{q\epsilon_0/d}$ is integrable (following from $\epsilon_0 \leq \epsilon$, and the fact that $f_{q\epsilon/d}$ is integrable by assumption), we can apply the dominated convergence theorem to yield

$$\begin{aligned}\lim_{\delta\to 0^+} \int_{\mathbb{R}^d} f_\delta(x)\,\mathrm{d}x &= \int_{\mathbb{R}^d} \lim_{\delta\to 0^+} f_\delta(x)\,\mathrm{d}x \\ &\leq \int_{\mathcal{K}\setminus B(0,\tau)} \lim_{\delta\to 0^+} \frac{|h(x)-h(0)|}{\delta^{d/q}} \cdot \exp(-\epsilon_0/\delta)\,\mathrm{d}x \\ &= 0,\end{aligned}$$

where the last line holds since $\delta^{-d/q}\exp(-\epsilon_0/\delta) \to 0$ as $\delta \to 0^+$.

**Step 4: analyzing the denominator in (50).** Recall that $x^\star = 0$, and we assume that the minimizer is in the interior of $\mathcal{K}$, thus we may assume without loss of generality that $B(0,\tau) \subseteq \mathcal{K}$ (otherwise in step 1 we simply take $\tau$ to be smaller). We may further assume that the local Hölder condition in (48) holds on $B(0,\tau)$ (again, otherwise in step 1 we just take $\tau$ to be smaller). We can then lower bound the denominator in (50) as follows:

$$\begin{aligned}\delta^{-d/q} \cdot \int_{\mathcal{K}\cap B(0,\tau)} \exp(-\varphi(x)/\delta)\,\mathrm{d}x &\geq \delta^{-d/q} \cdot \int_{B(0,\tau)} \exp(-a\|x\|_2^q/\delta)\,\mathrm{d}x \\ &= \int_{B(0,\tau\delta^{-1/q})} \exp(-a\|y\|_2^q)\,\mathrm{d}y \\ &= \int_0^{\tau\delta^{-1/q}} \int_{\partial B(0,r)} \exp(-ar^q)\,\mathrm{d}\mathcal{H}^{d-1}\,\mathrm{d}r,\end{aligned}$$

where the second line follows from a change of variables $x = \delta^{1/q} y$, and the last line from a change to polar coordinates (see, e.g., Appendix C.3 in Evans (2010)), with $\mathcal{H}^{d-1}$ denoting Hausdorff measure of dimension $d-1$. Now, $\mathcal{H}^{d-1}(\partial B(0,r)) = cr^{d-1}$ for a constant $c > 0$ depending only on $d$, so

$$\int_0^{\tau\delta^{-1/q}} \int_{\partial B(0,r)} \exp(-ar^q)\, \mathrm{d}\mathcal{H}^{d-1}\, \mathrm{d}r = c \cdot \int_0^{\tau\delta^{-1/q}} r^{d-1} \exp(-ar^q)\, \mathrm{d}r,$$

and therefore in the limit the denominator is lower bounded by

$$\lim_{\delta \to 0^+} c \cdot \int_0^{\tau\delta^{-1/q}} r^{d-1} \exp(-ar^q)\, \mathrm{d}r = c \cdot \int_0^{\infty} r^{d-1} \exp(-ar^q)\, \mathrm{d}r,$$

with the quantity in the last line being a positive constant. This completes the proof of the theorem. □

## A.3 Local Hölder condition for weakly convex functions

Recall, a function $\varphi$ is said to be $\rho$-weakly convex on a set $U$ provided that the map $x \mapsto \varphi(x) + (\rho/2)\|x\|_2^2$ is convex on $U$. We now verify that if $f$ and $g$ are $\rho_f$-weakly convex and $\rho_g$-weakly convex, respectively, on a neighborhood $U$ of $y_x$, then the local Hölder condition (25) in Corollary 1 holds for $\varphi_x$ with exponent $q = 1$. Note first that convexity of $f(y) + (\rho_f/2)\|y\|_2^2$ implies convexity of

$$f(y) - f(y_x) + \frac{\rho_f}{2}\|y\|_2^2 - 2y^\mathsf{T} y_x + \frac{\rho_f}{2}\|y_x\|_2^2 = \underbrace{f(y) - f(y_x) + \frac{\rho_f}{2}\|y - y_x\|_2^2}_{F_x(y)},$$

because we have only added a linear function and a constant. Similarly, convexity of $g(y) + (\rho_g/2)\|y\|_2^2$ implies convexity of

$$g(x-y) - g(x-y_x) + \frac{\rho_g}{2}\|x-y\|_2^2 - 2(x-y)^\mathsf{T}(x-y_x) + \frac{\rho_g}{2}\|x-y_x\|_2^2 = \underbrace{g(x-y) - g(x-y_x) + \frac{\rho_g}{2}\|y-y_x\|_2^2}_{G_x(y)},$$

because we have only made an affine variable transformation $y \mapsto x - y$, then added a linear function and a constant. Now, by definition

$$\begin{aligned}\varphi_x(y) - \varphi_x(y_x) &= f(y) - f(y_x) + g(x-y) - g(x-y_x) \\ &= F_x(y) + G_x(y) - (\rho_f + \rho_g)\|y - y_x\|_2^2.\end{aligned}$$

As $F_x(y) + G_x(y)$ is convex, and convex functions are locally Lipschitz (see, e.g., Theorem 6.7 in Evans and Gariepy (2015)), we know that it is Lipschitz on $U$. Furthermore, assuming without a loss of generality that $\mathrm{diam}(U) = \sup\{\|y - z\|_2 : y, z \in U\} \le 1$ (otherwise we just shrink $U$ around $y_x$ such that this holds), we have

$$(\rho_f + \rho_g)\|y - y_x\|_2^2 \le (\rho_f + \rho_g)\|y - y_x\|_2,$$

Thus $\varphi_x(y) - \varphi_x(y_x)$ is the sum of two Lipschitz functions on $U$, and therefore it is itself Lipschitz on $U$, i.e., locally Hölder with exponent $q = 1$.

## A.4 Proof of Theorem 3

For ease of reference, we state the theorem before its proof.

**Theorem 3.** *Under the conditions of Theorem 2, assume instead that the unique global minimizer $x^\star$ of $\varphi$ lies on the boundary of the closed set $\mathcal{K}$. Furthermore, assume that $\mathcal{K}$ is full-dimensional and star-shaped, in a local sense around $x^\star$: precisely, writing $B(x,r) = \{y \in \mathbb{R}^d : \|y - x\|_2 \le r\}$ for the ball of radius $r$ centered at $x$, we assume that there exists $r_0 > 0$ such that $\mathcal{K} \cap B(x^\star, r_0)$ has positive Lebesgue measure, and*

$$x \in \mathcal{K} \cap B(x^\star, r_0) \implies \alpha x + (1-\alpha)x^\star \in \mathcal{K}, \quad \text{for all } \alpha \in [0,1]. \tag{51}$$

*Then the same conclusion as in (49) holds.*

*Proof.* The proof is the same as that for Theorem 2 in Appendix A.2, except for step 4. Because $x^\star = 0$ is assumed to lie on the boundary of $\mathcal{K}$, it is no longer possible to take $\tau$ to be small enough so that $B(0,\tau) \subseteq \mathcal{K}$. However, we can still take $\tau$ to be small enough so that the local Hölder bound holds on $\mathcal{K} \cap B(0,\tau)$, which results in a lower bound for the denominator in (50) of

$$\begin{aligned}\delta^{-d/q} \cdot \int_{\mathcal{K}\cap B(0,\tau)} \exp(-\varphi(x)/\delta)\,\mathrm{d}x &\geq \delta^{-d/q} \cdot \int_{\mathcal{K}\cap B(0,\tau)} \exp(-a\|x\|_2^q/\delta)\,\mathrm{d}x \\ &= \int_{\mathcal{K}_\delta \cap B(0,\tau\delta^{-1/q})} \exp(-a\|y\|_2^q)\,\mathrm{d}y, \end{aligned} \tag{52}$$

the second line using a change of variables $x = \delta^{1/q} y$, where we abbreviate $\mathcal{K}_\delta = \delta^{-1/q}\mathcal{K}_\delta = \{\delta^{-1/q} x : x \in \mathcal{K}\}$. The rest of this proof proceeds as follows. We first show that we can use polar coordinates to compute (52), despite the fact that the integral is over the (possibly) nonspherical set $\mathcal{K}_\delta \cap B(0,\tau\delta^{-1/q})$ (step 4a). We then show how to lower bound this integral by a positive constant as $\delta \to 0^+$, using our assumptions on the set $\mathcal{K}$ (step 4b). This will complete the proof.

**Step 4a: rewriting (52) using polar coordinates.** Rewrite (52) as

$$\int_{\mathcal{K}_\delta\cap B(0,\tau\delta^{-1/q})} \exp(-a\|y\|_2^q)\,\mathrm{d}y = \int_{B(0,\tau\delta^{-1/q})} \underbrace{\mathbb{1}_{\mathcal{K}_\delta}(y) \cdot \exp(-a\|y\|_2^q)}_{\psi(y)}\,\mathrm{d}y.$$

Let $v_k = \eta_{1/k} * \mathbb{1}_{\mathcal{K}_\delta}$ be a mollified version of the indicator function $\mathbb{1}_{\mathcal{K}_\delta}$, where $\eta_\epsilon$ is the "standard" mollifier with bandwidth $\epsilon > 0$ (see, e.g., Appendix C.5 in Evans (2010)). Then defining $\psi_k(y) = v_k(y) \cdot \exp(-a\|y\|_2^q)$, since mollified functions converge locally in $L^1$ (e.g., Theorem 7 in Appendix C.5 of Evans (2010)),

$$\begin{aligned}\int_{B(0,\tau\delta^{-1/q})} \psi(y) &= \lim_{k\to\infty} \int_{B(0,\tau\delta^{-1/q})} \psi_k(y) \\ &= \lim_{k\to\infty} \int_0^{\tau\delta^{-1/q}} \int_{\partial B(0,r)} v_k \exp(-ar^q)\,\mathrm{d}\mathcal{H}^{d-1}\,\mathrm{d}r \\ &= \int_0^{\tau\delta^{-1/q}} \int_{\partial B(0,r)} \lim_{k\to\infty} v_k \exp(-ar^q)\,\mathrm{d}\mathcal{H}^{d-1}\,\mathrm{d}r \\ &= \int_0^{\tau\delta^{-1/q}} \int_{\partial B(0,r)} \mathbb{1}_{\mathcal{K}_\delta} \cdot \exp(-ar^q)\,\mathrm{d}\mathcal{H}^{d-1}\,\mathrm{d}r \\ &= \int_0^{\tau\delta^{-1/q}} \mathcal{H}^{d-1}(\mathcal{K}_\delta \cap \partial B(0,r)) \cdot \exp(-ar^q)\,\mathrm{d}r. \end{aligned} \tag{53}$$

The second line uses polar coordinates integration, which applies since $\psi_k$ is smooth (infinitely differentiable) by construction, the third uses the dominated convergence theorem, which applies because $0 \leq v_k \leq 1$, and the fourth uses the property that mollified functions converge pointwise almost everywhere (e.g., Theorem 7 in Appendix C.5 of Evans (2010)).

**Step 4b: lower bounding (53) as $\delta \to 0^+$.** Observe that the local star-shaped assumption (51) can be reformulated equivalently (recalling that we have taken $x^\star = 0$) as

$$\beta\mathcal{K} \cap \partial B(0,r) \supseteq \mathcal{K} \cap \partial B(0,r), \quad \text{for } r \leq r_0 \text{ and } \beta \geq 1. \tag{54}$$

To see this, take $x \in \mathcal{K}$ such that $\|x\|_2 = r \leq r_0$, and note that (51) says $\alpha x \in \mathcal{K}$ for $\alpha \leq 1$, or equivalently, $x \in \beta\mathcal{K}$ for $\beta = 1/\alpha \geq 1$. By taking $\delta \leq \min\{(\tau/r_0)^q, 1\}$, we have $\tau\delta^{-1/q} \geq r_0$, so we may lower bound the integral in (53) by

$$\int_0^{\tau\delta^{-1/q}} \mathcal{H}^{d-1}(\mathcal{K}_\delta \cap \partial B(0,r)) \cdot \exp(-ar^q)\,\mathrm{d}r \geq \int_0^{r_0} \mathcal{H}^{d-1}(\mathcal{K}_\delta \cap \partial B(0,r)) \cdot \exp(-ar^q)\,\mathrm{d}r.$$

Applying (54) to the integrand (with $\beta = \delta^{-1/q} \geq 1$) gives

$$\int_0^{r_0} \mathcal{H}^{d-1}(\mathcal{K}_\delta \cap \partial B(0,r)) \cdot \exp(-ar^q)\,\mathrm{d}r \geq \int_0^{r_0} \mathcal{H}^{d-1}(\mathcal{K} \cap \partial B(0,r)) \cdot \exp(-ar^q)\,\mathrm{d}r. \tag{55}$$

Meanwhile, for $r \leq r_0$,

$$\begin{aligned}\mathcal{H}^{d-1}(\mathcal{K} \cap \partial B(0,r)) &= \left(\frac{r}{r_0}\right)^{d-1} \mathcal{H}^{d-1}\left(\left(\frac{r_0}{r}\right)\mathcal{K} \cap \partial B(0,r_0)\right) \\ &\geq \left(\frac{r}{r_0}\right)^{d-1} \mathcal{H}^{d-1}(\mathcal{K} \cap \partial B(0,r_0)),\end{aligned} \tag{56}$$

the first line using the $(d-1)$-homogeneity of $\mathcal{H}^{d-1}$ (e.g., Theorem 2.2 in Chapter 2.1 of Evans and Gariepy (2015)), and the second line again using (54).

We remark our assumption $\mathcal{L}^d(\mathcal{K} \cap \partial B(0,r_0)) > 0$, where $\mathcal{L}^d$ denotes Lebesgue measure of dimension $d$, implies $\mathcal{H}^{d-1}(\mathcal{K} \cap \partial B(0,r_0')) > 0$, for some $r_0' \leq r_0$. Otherwise, $\mathcal{H}^{d-1}(\mathcal{K} \cap \partial B(0,r)) = 0$ for all $r \leq r_0$ would imply that $\mathcal{L}^d(\mathcal{K} \cap \partial B(0,r_0)) = 0$, as we can represent the Lebesgue measure as an integral over Hausdorff measure of boundary segments, by the same mollification argument used to derive (53) in step 4a. Assume without a loss of generality that $r_0' = r_0$ (otherwise we simply redefine $r_0$ to make this true), and abbreviate the lower bound in (56) by $cr^{d-1}$, where $c > 0$ is a constant depending only on $r_0$ and $d$. Then applying (56) to (55) gives

$$\int_0^{r_0} \mathcal{H}^{d-1}(\mathcal{K} \cap \partial B(0,r)) \cdot \exp(-ar^q)\,\mathrm{d}r \geq c \cdot \int_0^{r_0} r^{d-1} \exp(-ar^q)\,\mathrm{d}r,$$

the quantity in the last line being another positive constant. This completes the proof of the theorem. □

## A.5 Local full-dimensional and star-shaped conditions for convex bodies

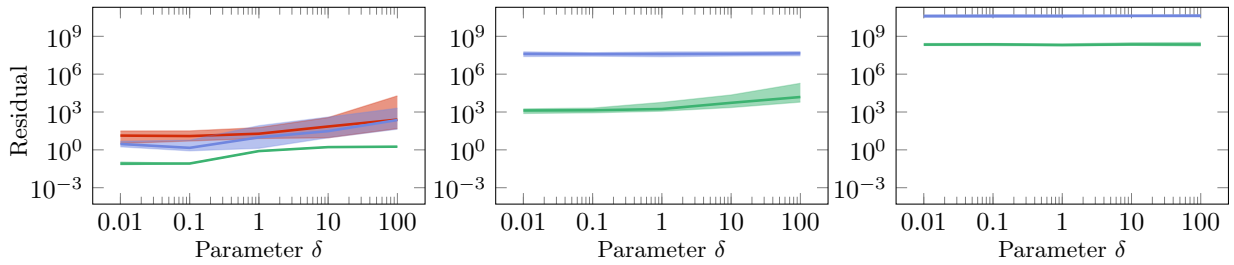
Recall, a convex body $\mathcal{K} \subseteq \mathbb{R}^d$ is a closed convex set with nonempty interior. Fix any $x^\star$ on the boundary of $\mathcal{K}$. For any $x \in \mathcal{K}$, we have $\alpha x + (1-\alpha)x^\star \in \mathcal{K}$ for all $\alpha \in [0,1]$ by convexity, so clearly the local star-shaped condition (26) is met for any $r_0 > 0$.

Meanwhile, for any $r_0 > 0$, the set $\mathcal{K} \cap B(x^\star, r_0)$ has nonempty interior. To see this, take $x \in \mathrm{int}(\mathcal{K})$ such that $\|x - x^\star\|_2 \leq r_0/3$. (This can be accomplished by taking an arbitrary point in the interior, then shrinking toward $x^\star$, and invoking convexity, until the distance bound is met.) Then by definition, there exists $\epsilon > 0$ such that $B(x,\epsilon) \subseteq \mathrm{int}(\mathcal{K})$. Defining $\epsilon' = \min\{\epsilon, r_0/3\}$, we have that $B(x,\epsilon')$ is contained in

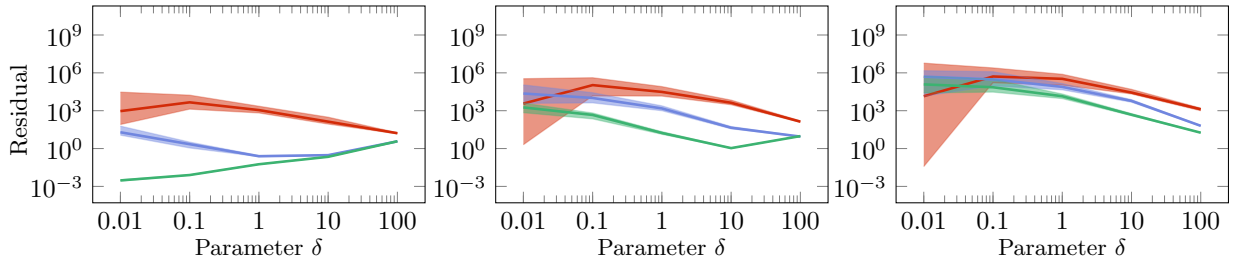
$$\mathrm{int}(\mathcal{K}) \cap \mathrm{int}(B(x^\star, r_0)) = \mathrm{int}(\mathcal{K} \cap B(x^\star, r_0)),$$

so $x$ lies in the interior of $\mathcal{K} \cap B(x^\star, r_0)$. A nonempty interior implies that $\mathcal{K} \cap B(x^\star, r_0)$ has positive Lebesgue measure. This verifies both conditions of Theorem 3 for convex bodies.

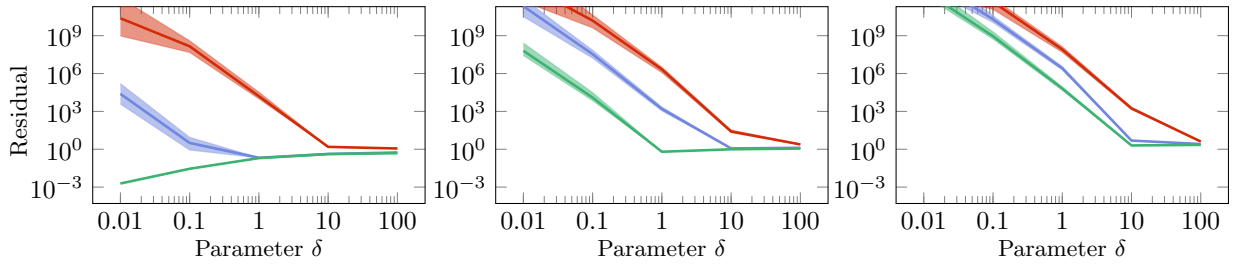
# B Further Hamilton-Jacobi Experiments



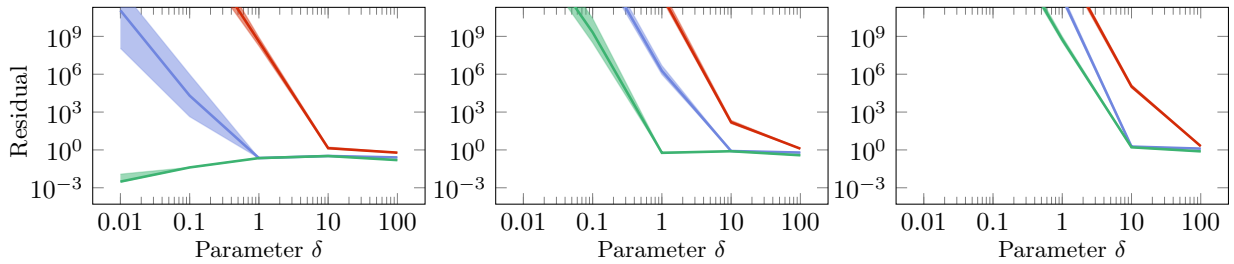

(a) Dimension $d = 2$, Norm $p = 1.1$ (b) Dimension $d = 5$, Norm $p = 1.1$ (c) Dimension $d = 10$, Norm $p = 1.1$




(d) Dimension $d = 2$, Norm $p = 2$ (e) Dimension $d = 5$, Norm $p = 2$ (f) Dimension $d = 10$, Norm $p = 2$




(g) Dimension $d = 2$, Norm $p = 5$ (h) Dimension $d = 5$, Norm $p = 5$ (i) Dimension $d = 10$, Norm $p = 5$




(j) Dimension $d = 2$, Norm $p = 10$ (k) Dimension $d = 5$, Norm $p = 10$ (l) Dimension $d = 10$, Norm $p = 10$

Figure 9: Residuals for HJ approximation in all dimensions $d$ and norms $p$ (for the Hamiltonian $H = \|\cdot\|_p^p/p$) that we consider. Each panel shows $N = 10$ samples via red; $N = 10^3$ samples via blue; and $N = 10^5$ samples via green. In general, we see similar trends to what is observed in Figure 5, and additionally, we now see that as the dimension grows, the errors also grow. The case $p = 1.1$ in dimensions $d = 5$ and $d = 10$ seems to be an exception, where we see poor accuracy even for large $N$ and small $\delta$ (with the residuals for $N = 10$ so large that they do not even appear in the plotting window). This may be due to the instability of autodiff in this case.