Non-Local Optimization: Imposing Structure on Optimization Problems by Relaxation

Nils Müller
Ruhr University Bochum
Bochum, Germany
mail@nilsmueller.io

Tobias Glasmachers
Institute for Neural Computation
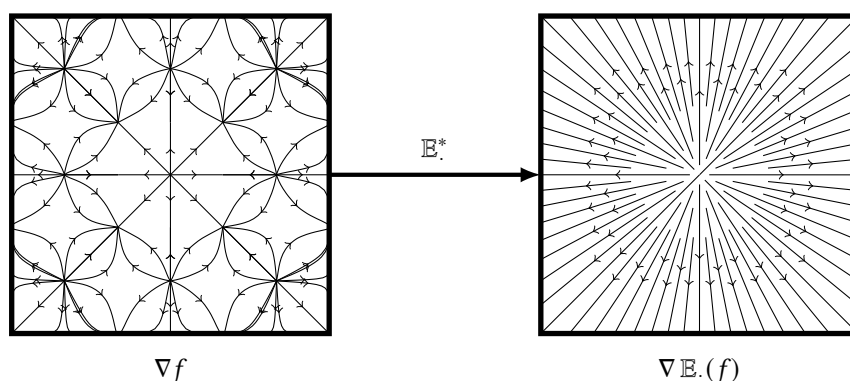Ruhr University Bochum
Bochum, Germany
tobias.glasmachers@ini.rub.de

∇f $\nabla \mathbb{E}_{\cdot}(f)$

Figure 1: Gradient flows of the Rastrigin function and its isotropic Gaussian relaxation.

ABSTRACT

In stochastic optimization, particularly in evolutionary computation and reinforcement learning, the optimization of a function $f : \Omega \to \mathbb{R}$ is often addressed through optimizing a so-called relaxation $\theta \in \Theta \mapsto \mathbb{E}_\theta(f)$ of f, where Θ resembles the parameters of a family of probability measures on Ω. We investigate the structure of such relaxations by means of measure theory and Fourier analysis, enabling us to shed light on the success of many associated stochastic optimization methods. The main structural traits we derive and that allow fast and reliable optimization of relaxations are the consistency of optimal values of f, Lipschitzness of gradients, and convexity. We emphasize settings where f itself is not differentiable or convex, e.g., in the presence of (stochastic) disturbance.

CCS CONCEPTS

• **Mathematics of computing → Mathematical optimization**;
• **Theory of computation → Theory of randomized search heuristics**.

KEYWORDS

Global optimization, Stochastic optimization, Evolution strategies, Robust optimization

1 INTRODUCTION

The idea of an optimization problem is to find minima (or maxima) of a function $f : \Omega \to \mathbb{R}$ in a set of possible candidates Ω. Given a σ-algebra $\mathcal{A}$ on Ω we can understand a probability measure Pr on $(\Omega, \mathcal{A})$ as a generalized candidate of f. A generalized candidate Pr induces a functional to which we assign its expected value $\mathbb{E}(f)$. A parameterization $\text{Pr}_{\cdot} : \Theta \to \mathcal{M}_1^+(\Omega, \mathcal{A})$ of (a subset of) the generalized candidates, where Θ is the parameter set and $\mathcal{M}_1^+(\Omega, \mathcal{A})$ is the set of probability measures on $(\Omega, \mathcal{A})$, results in what is called a **stochastic relaxation** $\theta \in \Theta \mapsto \mathbb{E}_\theta(f)$. In this paper, we investigate the central structural properties of relaxed problems that are of interest to optimization: consistency, smoothness, and convexity.

Algorithms. Our work applies to a broad range of optimization algorithms, including several randomized search heuristics (variants of estimation of distribution methods) for discrete and continuous search spaces. The second half of our paper is heavily influenced by applications in gradient-based optimization, covering in particular information geometric optimization [12]. The two most important

use-cases in the regime of gradient-based optimization are the following:

- Θ is an open subset of $\mathbb{R}^n$ and the measures have θ-differentiable densities with respect some measure μ, i.e. $\Pr_\theta = k_\theta \mu$.
- $\Omega = \Theta = \mathbb{R}^n$, $\mathcal{A}$ is the Borel σ-algebra of $\mathbb{R}^n$, i.e., $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$, f is differentiable, and $\Pr_\theta \circ \psi_\theta = \Pr_0$, where $\psi_\theta : \mathbb{R}^n \to \mathbb{R}^n$, $x \mapsto x + \theta$.

Under mild regularity assumptions, which we will cover, we have the identity

$$\nabla_\theta \, \mathbb{E}.(f) \equiv \mathbb{E}.(f \nabla_\theta \ln k.)$$

in the first case and

$$\nabla_\theta \, \mathbb{E}.(f) \equiv \mathbb{E}.(\nabla_x f)$$

in the second. In both cases, the gradient of the relaxation can, therefore, be approximated by numerical integration.

Motivation and Related Work. Stochastic relaxations have long provided a powerful approach to optimization problems where non-local structure is significant, problem representations are not easy to manipulate (e.g., through differentiation), or robust solutions are desired. In recent literature, there has been considerable progress on qualitative and quantitative assessment of optimization methods of such relaxations' parameter spaces, particularly concerning their convergence on problems where local structure is instructive [1, 6, 11, 16]. Even before, there have been numerous proposals of frameworks that guide the principled design of optimization methods for a wide range of discrete and continuous problems [7, 12, 15]. The use of stochastic relaxations, however, has been motivated by invariance of solutions under transformations [12], biological plausibility of evolutionary computation [15], program simplicity, as well as practical utility [11, 13], or by experiment [3, 13]. Yet the favorable structure of stochastic relaxations for optimization itself has received a rather incidental treatment in favor of different questions. Outstanding from the rest of the literature, [4, 9, 10] derive insightful and principled properties of special relaxations under decay and convexity conditions that arguably deviate from the classical setting in optimization. Furthermore, deep connections of Gaussian relaxations and approximation have been established in [8]. Surprisingly, much of the referenced prior work seems to have been compiled in virtually independent communities.

In this work, we investigate the relation between a problem and its stochastic relaxations and derive a systematic understanding of why stochastic relaxations are favorable for fast and reliable optimization in practice.

Outline. In Section 2, we showcase the structure of a stochastic relaxation of the Rastrigin function, a popular highly multi-modal benchmark problem in optimization. We also introduce the general problem definition covering discrete and continuous spaces. In Section 3, a criterion for the consistency of optimal function values and the location of optima is established. In Section 4, we derive representations of derivatives of stochastic relaxations. Among this we cover the transfer of (Lipschitz) differentiability from parameterized densities to the stochastic relaxation. Here we tie together many notions of differentiability that are used in practice. In Section 5, we derive insightful sufficient conditions under which functions on $\mathbb{R}^n$ have a convex stochastic relaxation based on a notion of weak convexity and Fourier analysis.

2 TRANSFORMING AN OPTIMIZATION PROBLEM

In this section, we will demonstrate by example, what we will establish in later sections for more general cases and what can be understood as the driving factors for the success of stochastic relaxations. We chose a popular benchmark problem from optimization literature, the Rastrigin function, and multivariate isotropic normal distributions as parameterized probability measures. This setting is typical in stochastic optimization, particularly with evolution strategies. More concretely, we want to illustrate the following traits of this particular stochastic relaxation:

- Consistency of the minimum of the original problem,
- Lipschitz continuity of gradients, and
- convexity.

The above traits will serve to guide our intuition in the subsequent sections.

EXAMPLE 2.1. *Consider the Rastrigin function*

$$x \in \mathbb{R}^n \longmapsto f(x) := \|x\|^2 - \sum_{j=1}^{n} a_j \cos(\xi_j x_j),$$

where $\xi_j, a_j > 0$ for all $j \in \mathbb{N}_{\leq n}$. Let $\Pr_{\theta,\sigma} \sim \mathcal{N}(\theta, \sigma^2 I)$, where $\theta \in \mathbb{R}^n$ and $\sigma > 0$. Rastrigin's function is an example of a highly multi-modal optimization problem. Its global minimum is found at the origin, while an exponentially large (in the problem dimension) number of local minima represent distractors to local optimization methods.

We will now look at the structure of the function

$$\theta \in \mathbb{R}^n \longmapsto \mathbb{E}_{\theta,\sigma}(f),$$

which we interpret as a relaxation on the original candidate space $\mathbb{R}^n$ of f. We find the following traits.

- *We can pick a small σ such that we can approximate the value of $f(x)$ with $\mathbb{E}_{x,\sigma}(f)$ with arbitrary precision—in particular, at the global minimum of f at 0.*
- *Structural traits such as the smoothness of the density of the normal distribution are transferred to $\mathbb{E}.(f)$. Moreover, the gradients of $\theta \mapsto \mathbb{E}_{\theta,\sigma}(f)$ are Lipschitz for all $\sigma > 0$.*
- *As the following arguments will show, we can pick σ^* such that $\theta \mapsto \mathbb{E}_{\theta,\sigma}(f)$ is strictly convex for all $\sigma > \sigma^*$.*

In our specific example we obtain the traits in terms of handy closed-form formulas:

(1) Let λ denote the Lebesgue measure of $\mathbb{R}^n$. For all $\theta \in \mathbb{R}^n$ and $\sigma > 0$, we have

$$\mathbb{E}_{\theta,\sigma}(\|\cdot\|^2) = \int_{\mathbb{R}^n} \sum_{k=1}^{n} x_k^2 \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\|\theta - x\|^2\right) \lambda(\mathrm{d}x)$$

(definitions)

$$= \sum_{k=1}^{n} \int_{\mathbb{R}^n} x_k^2 \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\|\theta - x\|^2\right) \lambda(\mathrm{d}x)$$

(linearity)

$$= \sum_{k=1}^{n} \int_{\mathbb{R}} x_k^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\theta_k - x_k)^2\right) \lambda(\mathrm{d}x_k)$$

(Fubini, linearity, and probability measure)

$$= n\sigma^2 + \|\theta\|^2 \, .$$

(Steiner's Theorem)

(2) Let $j \in \mathbb{N}_{\leq n}$ and define $\Xi_j := \xi_j e_j$, then, using the representation of the characteristic function of the multivariate normal distribution in the third equality, we obtain

$$\mathbb{E}_{\theta,\sigma}(\cos(\xi_j(\cdot)_j)) = \int_{\mathbb{R}^n} \mathrm{Re}\left(e^{i\langle \Xi_j, x\rangle}\right) \mathrm{Pr}_{\theta,\sigma}(\mathrm{d}x)$$

(definitions)

$$= \mathrm{Re}\left(\int_{\mathbb{R}^n} e^{i\langle \Xi_j, x\rangle} \mathrm{Pr}_{\theta,\sigma}(\mathrm{d}x)\right)$$

(definition of the complex integral)

$$= \mathrm{Re}\left(\exp\left(i\langle \Xi_j, \theta\rangle - \frac{1}{2}\sigma^2\|\Xi_j\|^2\right)\right)$$

(characteristic function of the multivariate normal)

$$= \mathrm{Re}\left(\exp(i\xi_j\theta_j)\right) \exp\left(-\frac{1}{2}\sigma^2\xi_j^2\right)$$

(definitions and exponential rule)

$$= \cos(\xi_j\theta_j) \exp\left(-\frac{1}{2}\sigma^2\xi_j^2\right) \, .$$ (definition of cos)

Therefore, by linearity of the integral, we have

$$\mathbb{E}_{\theta,\sigma}(f) = n\sigma^2 + \|\theta\|^2 - \sum_{j=1}^{n} a_j \cos(\xi_j\theta_j) \exp\left(-\frac{1}{2}\sigma^2\xi_j^2\right) \, .$$

Now, we can check strict convexity by looking at the Hessian. The second partial derivatives of the relaxation are given by

$$\partial_{\theta_j,\theta_k} \mathbb{E}_{\theta,\sigma}(f) = \delta_{jk} \cdot \left(2 + a_j\xi_j^2 \cos(\xi_j\theta_j) \exp\left(-\frac{1}{2}\sigma^2\xi_j^2\right)\right),$$

where δ_{jk} is the Kronecker delta. Thus, the Hessian is diagonal with strictly positive values at $\theta \in \mathbb{R}^n$ if and only if for all $j \in \mathbb{N}_{\leq n}$

$$-2 < a_j\xi_j^2 \cos(\xi_j\theta_j) \exp\left(-\frac{1}{2}\sigma^2\xi_j^2\right) \, .$$

As $\cos(\xi_j\theta_j) \geq -1$, we know that this is the case at least if

$$2 > a_j\xi_j^2 \exp\left(-\frac{1}{2}\sigma^2\xi_j^2\right) \iff \frac{2}{\xi_j^2}\log\left(\frac{a_j\xi_j^2}{2}\right) < \sigma^2 \, .$$

Picking $\sigma^{*2} > \max_{j\in\mathbb{N}_{\leq n}} \frac{2}{\xi_j^2}\log\left(\frac{a_j\xi_j^2}{2}\right)$, we obtain the result.

The above result outlines the very structure that can be exploited for optimization by first-order optimization methods [2]. In some practical settings, a gradient method on the mean parameter θ may be augmented by a manipulation of the standard deviation σ. However, keeping σ large improves robustness.

What follows is the general setting used throughout the work. To this end, let $\mathcal{L}^1(\Omega, \mathcal{A}, \mathrm{Pr}_\theta)$ denote the at least once Lebesgue integrable, real-valued functions on the measure space $(\Omega, \mathcal{A}, \mathrm{Pr}_\theta)$.

Definition 2.1 (Stochastic Relaxation). *Given*

- *a family of probability measures* $\{\mathrm{Pr}_\theta : \theta \in \Theta\}$ *on a measure space* $(\Omega, \mathcal{A})$,
- *an optimization problem* $\min_{x\in\Omega} f(x)$, *where* $f : \Omega \to \mathbb{R}$, *and*
- $f \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathrm{Pr}_\theta)$, $\forall \theta \in \Theta$,

we call the problem

$$\min_{\theta\in\Theta} \mathbb{E}_\theta(f) = \min_{\theta\in\Theta} \int_\Omega f \, \mathrm{dPr}_\theta \, ,$$

the **stochastic relaxation** *of f for which we write* $(f, \Omega, \mathcal{A}, (\mathrm{Pr}_\theta)_{\theta\in\Theta})$. *In case Ω has a metric d such that $\mathcal{B}(d) = \mathcal{A}$ we write* $(f, d, \Omega, \mathcal{A}, (\mathrm{Pr}_\theta)_{\theta\in\Theta})$.[1]

Remark 2.1. *In principle, there is no need for requiring the measures $(\mathrm{Pr}_\theta)_{\theta\in\Theta}$ to be non-negative. One could define a relaxed problem analogously by allowing signed measures. In this case however, the relation of the relaxed problem and the original problem in Section 3 or the results of Section 5 will in general not hold.*

Inspired by the previous example, we will now aim to generalize the traits of Example 2.1 to fit the picture of the practical use of stochastic relaxations.

3 CONSISTENCY IN THE CODOMAIN

Manipulating an optimization problem raises the question of whether solutions or the cost of solutions will relate to those of the original optimization problem defined by a function $f : \Omega \to \mathbb{R}$.

We can observe that some distinguished generalized candidates can be associated with a set of close-to-optimal candidates of f by their function value and mass distribution in candidate space. Formally, we define the following.

Definition 3.1 (f-consistent). *Let $(f, d, \Omega, \mathcal{A}, (\mathrm{Pr}_\theta)_{\theta\in\Theta})$ be a relaxation as defined in Definition 2.1 and let $f : \Omega \to \mathbb{R}$ have a unique global minimum at x^*. We call the relaxation f-**consistent** if*

$$\forall x \in \Omega, \ x \neq x^* : \exists \theta^* \in \Theta : \ \mathbb{E}_{\theta^*}(f) < f(x) \, .$$

We will learn that requiring the Dirac measure at the global optimum of f to be a limit candidate of the relaxation in conjunction with a bounding attribute of f is sufficient for a relaxation to be f-consistent. A concise way of treating the problem based on measure theory is presented in this section. The results cover a general setting that includes the candidate spaces $\Omega \in \{\mathbb{Z}^n, \mathbb{R}^n\}$.

Definition 3.2 (f-ε-δ-concentrated, f-concentrated). *Let $(f, d, \Omega, \mathcal{A}, (\mathrm{Pr}_\theta)_{\theta\in\Theta})$ be a relaxation as defined in Definition 2.1. If $\varepsilon, \delta > 0$, the measure Pr_θ is called f-ε-δ-**concentrated** at $x \in \Omega$ if*

$$\int_{\Omega-U_\delta(x)} \max\{|f|, 1\} \, \mathrm{dPr}_\theta < \varepsilon, \quad \text{(CON)}$$

*where $U_\delta(x)$ denotes the δ-ball centered at x. Moreover, probability measures $(\mathrm{Pr}_\theta)_{\theta\in\Theta}$ are called f-**concentrated** at $x \in \Omega$ if*

$$\forall \varepsilon, \delta > 0 : \exists \theta \in \Theta : \ \mathrm{Pr}_\theta \ \text{is } f\text{-}\varepsilon\text{-}\delta \ \text{concentrated at } x \, .$$

A general approximation result will pave the way toward approximation at a global optimum of f.

---

[1]We denote the Borel σ-algebra of a topological space Ω by $\mathcal{B}(\Omega)$. If Ω admits a metric d, we also write $\mathcal{B}(d)$.

LEMMA 3.1 (APPROXIMATION). *Let $(f, d, \Omega, \mathcal{A}, (\Pr_\theta)_{\theta \in \Theta})$ be a relaxation as defined in Definition 2.1, where $(\Pr_\theta)_{\theta \in \Theta}$ are f-concentrated at $x \in \Omega$ and f is continuous at the same $x \in \Omega$. We have*

$$\forall \gamma > 0 : \exists \theta \in \Theta : |\mathbb{E}_\theta(f) - f(x)| < \gamma.$$

PROOF. Let $\gamma > 0$. Due to continuity of f at $x \in \Omega$, we pick $\delta > 0$ such that

$$d(x, y) < \delta \implies |f(x) - f(y)| < \frac{\gamma}{3}. \tag{1}$$

As $(\Pr_\theta)_{\theta \in \Theta}$ are f-concentrated at x, we pick θ^* such that

$$\Pr_{\theta^*} \text{ is } f\text{-}\varepsilon\text{-}\delta\text{-concentrated at } x, \tag{2}$$

where $\varepsilon := \min\left\{\frac{\gamma}{3}, \frac{\gamma}{3|f(x)|}\right\}$.

It follows, that

$$|\mathbb{E}_{\theta^*}(f) - f(x)|$$
$$= \left| \int_{\Omega - U_\delta(x)} f \, d\Pr_{\theta^*} + \int_{U_\delta(x)} f \, d\Pr_{\theta^*} - f(x) \right| \quad \text{(additivity)}$$
$$\leq \int_{\Omega - U_\delta(x)} |f| \, d\Pr_{\theta^*} + \left| \int_{U_\delta(x)} f \, d\Pr_{\theta^*} - f(x) \right|$$
$$\text{(triangle inequality)}$$
$$< \frac{\gamma}{3} + \left| \int_{U_\delta(x)} f \, d\Pr_{\theta^*} - f(x) \right| \quad \text{(Equation (2))}$$
$$= \frac{\gamma}{3} + \left| \int_{U_\delta(x)} f - f(x) \, d\Pr_{\theta^*} + f(x)\Pr_{\theta^*}(\Omega - U_\delta(x)) \right|$$
$$\text{(additivity)}$$
$$\leq \frac{\gamma}{3} + \int_{U_\delta(x)} |f - f(x)| \, d\Pr_{\theta^*} + |f(x)| \Pr_{\theta^*}(\Omega - U_\delta(x))$$
$$\text{(triangle inequality)}$$
$$\leq \frac{\gamma}{3} + \frac{\gamma}{3}\Pr_{\theta^*}(U_\delta(x)) + |f(x)| \Pr_{\theta^*}(\Omega - U_\delta(x)) \quad \text{(Equation (1))}$$
$$< \frac{\gamma}{3} + \frac{\gamma}{3} + \frac{\gamma}{3} = \gamma \quad \text{(probability measure and Equation (2))}$$

$\square$

REMARK 3.1. *In general, we can not relax the requirement of f-concentration as specified by (CON) and still fulfill the approximation property given in Lemma 3.1. Consider the following examples.*

(1) Let $f : \mathbb{R} \to \mathbb{R}, x \mapsto x^2$, δ_x denote the Dirac measure at $x \in \mathbb{R}$, and

$$(\Omega, \mathcal{A}, \Pr_\theta) \overset{!}{=} (\mathbb{R}, \mathcal{B}(\mathbb{R}), (1-\theta)\delta_0 + \theta\delta_{1/\theta}), \ \forall \theta \in (0,1).$$

Clearly, we have a concentration of measure at $x = 0$, as

$$\Pr_\theta(\Omega - \{0\}) = \theta, \ \forall \theta \in (0,1).$$

However, we can not approximate f at 0 as in Lemma 3.1, since for all $\theta \in (0,1)$, we have

$$\mathbb{E}_\theta(f) = \frac{1}{\theta} > 1 > 0 = f(0).$$

(2) Let $f : \mathbb{R} \to \mathbb{R}, x \mapsto -\exp(-x^2)$ and

$$(\Omega, \mathcal{A}, \Pr_\theta) \overset{!}{=} (\mathbb{R}, \mathcal{B}(\mathbb{R}), \delta_\theta), \ \forall \theta \in (1, \infty).$$

Clearly, the L^1-norm of our function outside of 0 can be arbitrarily small, i.e.,

$$\forall \varepsilon > 0 : \exists \theta \in (1, \infty) : \int_{\Omega - \{0\}} |f| \, d\delta_\theta \leq \exp(-\theta^2) < \varepsilon.$$

However, we can not approximate f at 0 as in Lemma 3.1, since for all $\theta \in (1, \infty)$, we have

$$\mathbb{E}_\theta(f) = -\exp(-\theta^2) > -\frac{1}{e} > -1 = f(0).$$

THEOREM 3.1 (CONSISTENCY). *Let $(f, d, \Omega, \mathcal{A}, (\Pr_\theta)_{\theta \in \Theta})$ be a relaxation as defined in Definition 2.1 and let $f : \Omega \to \mathbb{R}$ be continuous at its unique global minimum at x^*. Furthermore, assume that $(\Pr_\theta)_{\theta \in \Theta}$ are f-concentrated at x^*. Then, the relaxation is f-consistent.*

PROOF. Pick $x \in \Omega$ such that $x \neq x^*$. Define

$$\gamma := f(x) - f(x^*), \tag{3}$$

which is strictly positive as x^* is the unique global minimum of f.

By assumption, $(\Pr_\theta)_{\theta \in \Theta}$ are f-concentrated at x^*. Thus, by Lemma 3.1 we find $\theta^* \in \Omega$ such that

$$\left| \mathbb{E}_{\theta^*}(f) - f(x^*) \right| < \gamma. \tag{4}$$

Applying Equation (3) in the first step, Equation (4) in the second step and the fact that x^* is a global minimum of f in the third, we obtain the result

$$f(x) = \gamma + f(x^*)$$
$$> \left| \mathbb{E}_{\theta^*}(f) - f(x^*) \right| + f(x^*)$$
$$= \mathbb{E}_{\theta^*}(f) - f(x^*) + f(x^*)$$
$$= \mathbb{E}_{\theta^*}(f).$$

$\square$

A consistent relaxation ensures that minimizing the relaxation results in function values arbitrarily close to those of the f at its optimum $x^* \in \Omega$. Therefore, we find $\Pr_\theta$ with an arbitrarily close-to-one probability of an arbitrarily low regret. In the limiting case, we can also observe the following immediate result.

COROLLARY 3.1.1. *In the setting of Theorem 3.1, if the relaxation has a minimum, it is unique at the Dirac measure of the minimum of f, that is, at δ_{x^*}.*

4 IMPOSING DIFFERENTIABILITY

In this section, we derive gradient representations of relaxations. The results in Theorem 4.2 and Corollary 4.2.1 cover cases where the objective function transformed by the relaxation may not be differentiable or no differentiable structure of the candidate space Ω is specified.

First, however, let us look at the simple case where the candidate space $\Omega = \Theta = \mathbb{R}^n$, the objective function $f : \Omega \to \mathbb{R}$ is continuously differentiable, and there is a distinguished probability measure $\Pr$ on $(\Omega, \mathcal{B}(\Omega))$. We can construct a family of probability measures by setting

$$\Pr_\theta(A) := \Pr(A - \theta)$$

for all $\theta \in \Theta$ and for all $A \in \mathcal{A} := \mathcal{B}(\Omega)$. This case is critical in practice and particularly interesting due to the following property.

THEOREM 4.1 (PRESERVING DIFFERENTIABILITY). *Let a relaxation as defined in Definition 2.1 be given by $(f, \mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\Pr_\theta)_{\theta \in \mathbb{R}^n})$ with*

- $\Pr_\theta(A) = \Pr_0(A - \theta)$ *for all $\theta \in \mathbb{R}^n$ and for all $A \in \mathcal{B}(\mathbb{R}^n)$*

- $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, and
- there exists $\varepsilon > 0$ such that $x \in \mathbb{R}^n \mapsto \max |(\partial_r f)(U_\varepsilon(x))|$ has a $\Pr_0$-integrable upper bound for all $r \in \mathbb{N}_{\leq n}$.

We have $\nabla_\theta \mathbb{E}.(f) \equiv \mathbb{E}.(\nabla_x f)$. If $\nabla_x f$ is Lipschitz/uniformly continuous, then $\nabla_\theta \mathbb{E}.(f)$ is as well.

REMARK 4.1. *The third assumption deviates from the standard assumption of Leibniz's integral rule, as we have to restrict the parameter set to some neighborhood around any parameter to obtain a bound that fits Leibniz's integral rule. As differentiability is a local property, this still suffices for the statement of the theorem.*

PROOF. Let $\psi(\theta, x) := \psi_\theta(x) := x + \theta$ for all $x \in \mathbb{R}^n$ and for all $\theta \in \mathbb{R}^n$, then

$$\int_\Omega f \, \mathrm{d}\Pr_\theta = \int_0^\infty \Pr_\theta(\{x \in \mathbb{R}^n \mid f(x) > t\}) \, \mathrm{d}t$$

(definition Lebesgue integral)

$$= \int_0^\infty \Pr_0\big(\psi_\theta^{-1}(\{x \in \mathbb{R}^n \mid f(x) > t\})\big) \, \mathrm{d}t \quad \text{(assumption)}$$

$$= \int_0^\infty \Pr_0(\{x \in \mathbb{R}^n \mid (f \circ \psi_\theta)(x) > t\}) \, \mathrm{d}t$$

(definition inverse)

$$= \int_\Omega f \circ \psi_\theta \, \mathrm{d}\Pr_0 \,. \qquad \text{(definition Lebesgue integral)}$$

Due to the last assumption, we can apply Leibniz's integral rule [14, p. 91, Theorem 11.5], and get

$$\nabla_\theta \int_\Omega f \circ \psi_\theta \, \mathrm{d}\Pr_0 = \int_\Omega (\nabla_\theta f \circ \psi)(\theta, x) \, \Pr_0(\mathrm{d}x)$$

(Leibniz integral rule)

$$= \int_\Omega (\nabla_x f)(x + \theta) \, \Pr_0(\mathrm{d}x) \qquad \text{(definition } \psi)$$

$$= \int_\Omega (\nabla_x f) \, \mathrm{d}\Pr_\theta \,. \qquad \text{(above equation)}$$

Combining the results, we obtain for all $\theta \in \mathbb{R}^n$, that

$$\nabla_\theta \mathbb{E}_\theta(f) = \mathbb{E}_\theta(\nabla_x f) \,.$$

In case the gradient of f is also Lipschitz with constant L, we have

$$\|\nabla_\theta \mathbb{E}_{\theta'}(f) - \nabla_\theta \mathbb{E}_\theta(f)\|$$

$$\leq \int_\Omega \big\|(\nabla_x f)(x + \theta') - (\nabla_x f)(x + \theta)\big\| \, \Pr_0(\mathrm{d}x)$$

(above derivation)

$$\leq \int_\Omega L \, \|\theta' - \theta\| \, \Pr_0(\mathrm{d}x) \qquad \text{(Lipschitzness of } f)$$

$$= L \, \|\theta' - \theta\| \,. \qquad \text{(Lebesgue integral \& probability measure)}$$

By an analogous derivation, we obtain continuity of $\nabla_\theta \mathbb{E}.(f)$ if $\nabla_x f$ is continuous. □

We will now look again at the abstract, general setting for relaxations. A transfer of smoothness from the probability measures $\{\Pr_\theta \mid \theta \in \Theta\}$ to $\mathbb{E}.(f)$ is also central to applications. Tying together many notions of differentiability, e.g., for parameter-free distributions, we consider pullbacks using $\gamma : U \subseteq \mathbb{R}^n \to \Theta$ to describe the transfer of (Lipschitz) differentiablity.

THEOREM 4.2. *Let $(f, \Omega, \mathcal{A}, (\Pr_\theta)_{\theta \in \Theta})$ be a relaxation as defined in Definition 2.1 and let $\gamma : U \to \Theta$ be a function for some open $U \subseteq \mathbb{R}^n$.*

(Partial Differentiability) If
- *there exists a measure μ on $(\Omega, \mathcal{A})$ s.t. for all $t \in U$ there exists a density $k_{\gamma(t)}$ of $\Pr_{\gamma(t)}$ w.r.t. μ,*
- *for all $x \in \Omega$ the partial derivatives of $t \in U \mapsto k_{\gamma(t)}(x)$ exist, and*
- *for all $r \in \mathbb{N}_{\leq n}$ there exists $g_r \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ s.t. for all $(t, x) \in U \times \Omega$ we have*

$$\left| f(x) \partial_r k_{\gamma(t)}(x) \right| \leq g_r(x) \,, {}^2$$

then the partial derivatives of $t \in U \mapsto \mathbb{E}_{\gamma(t)}(f)$ exist and

$$\partial_r \mathbb{E}_{\gamma(t)}(f) = \int_\Omega f \partial_r k_{\gamma(t)} \, \mathrm{d}\mu \,.$$

(Continuous Differentiability) If in addtion, for all $r \in \mathbb{N}_{\leq n}$ and for all $x \in \Omega$, $t \in U \mapsto \partial_r k_{\gamma(t)}(x)$ is continuous, then $t \in U \mapsto \mathbb{E}_{\gamma(t)}(f)$ is continuously differentiable.
(Lipschitz Differentiability) If in addition, for all $r \in \mathbb{N}_{\leq n}$ and for all $x \in \Omega$, there exists $L_r(x) \in \mathbb{R}$ s.t. for all $t', t \in U$

$$\left| \partial_r k_{\gamma(t')}(x) - \partial_r k_{\gamma(t)}(x) \right| \leq L_r(x) \, \|t' - t\|$$

$$\text{and } f L_r \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu) \,,$$

then $t \in U \mapsto \nabla \mathbb{E}_{\gamma(t)}(f)$ is Lipschitz with constant

$$\sum_{r=1}^n \int_\Omega |f| \, L_r \, \mathrm{d}\mu \,.$$

PROOF.

(Partial Differentiability) To obtain the first result, we apply Leibniz's integral rule [14, p. 91, Theorem 11.5]. The requirements are fulfilled as per the following arguments.

(a) For all $t \in U$, we have

$$x \in \Omega \mapsto f(x) k_{\gamma(t)}(x) \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$$

as $f \in \mathcal{L}^1(\Omega, \mathcal{A}, \Pr_{\gamma(t)})$ and $k_{\gamma(t)} \mu = \Pr_{\gamma(t)}$ μ-almost everywhere.

(b) For all $x \in \Omega$, the function $t \in U \mapsto f(x) k_{\gamma(t)}(x)$ is partially differentiable as $t \in U \mapsto k_{\gamma(t)}(x)$ is partially differentiable.

(c) In each coordinate this requirement is directly fulfilled by our initial assumptions.

As U is open in $\mathbb{R}^d$, we can apply Leibniz's integral rule to each coordinate direction.

(Continuous Differentiability) To obtain the second result, we apply the Lebesgue dominated convergence theorem, e.g., see [14, p. 89], to an arbitrary convergent sequence $(t_j)_{j \in \mathbb{N}} \to t \in U$. Clearly, for all $r \in \mathbb{N}_{\leq n}$ and for all $x \in \Omega$

$$\partial_r k_{\gamma(t_j)}(x) \xrightarrow{j \to \infty} \partial_r k_{\gamma(t)}(x)$$

$$\implies \quad f(x) \partial_r k_{\gamma(t_j)}(x) \xrightarrow{j \to \infty} f(x) \partial_r k_{\gamma(t)}(x) \,.$$

The left side of the above statement is fulfilled by our additional continuity assumption. The boundedness requirement

---

^{2}The operator ∂_r refers to the partial derivative with respect to the r-th argument of the function γ.

is fulfilled by the assumption of the first statement. Therefore, the Lebesgue dominated convergence theorem tells us for all $r \in \mathbb{N}_{\le n}$

$$\lim_{j \to \infty} \int_\Omega f \partial_r k_{\gamma(t_j)} \, \mathrm{d}\mu = \int_\Omega f \partial_r k_{\gamma(t)} \, \mathrm{d}\mu.$$

Combined with the first result we get that for all $r \in \mathbb{N}_{\le n}$

$$\lim_{j \to \infty} \partial_r \mathbb{E}_{\gamma(t_j)}(f) = \partial_r \mathbb{E}_{\gamma(t)}(f),$$

i.e., continuity of the partial differentials. This implies continuity of the differential.

(Lipschitz Differentiability) To obtain the third result, we observe that for all $r \in \mathbb{N}_{\le n}$, for all $x \in \Omega$, and for all $t', t \in U$

$$\left| \partial_r k_{\gamma(t')}(x) - \partial_r k_{\gamma(t)}(x) \right| \le L_r(x) \left\| t' - t \right\|$$

$$\implies \quad \left| f(x) \partial_r k_{\gamma(t')}(x) - f(x) \partial_r k_{\gamma(t)}(x) \right| \le |f(x)| L_r(x) \left\| t' - t \right\|.$$

The left side of the above statement is fulfilled by our additional Lipschitz assumption. Using the result of the first statement and the triangle inequality this implies that for all $t', t \in U$

$$\left\| \nabla \mathbb{E}_{\gamma(t')}(f) - \nabla \mathbb{E}_{\gamma(t)}(f) \right\| = \left\| \int_\Omega f \nabla k_{\gamma(t')} - f \nabla k_{\gamma(t)} \, \mathrm{d}\mu \right\|$$

$$\le \int_\Omega \left\| f \nabla k_{\gamma(t')} - f \nabla k_{\gamma(t)} \right\| \, \mathrm{d}\mu$$

$$\le \int_\Omega \sum_{r=1}^n \left| f \partial_r k_{\gamma(t')} - f \partial_r k_{\gamma(t)} \right| \, \mathrm{d}\mu$$

$$\le \int_\Omega \sum_{r=1}^n |f| L_r \left\| t' - t \right\| \, \mathrm{d}\mu$$

$$= \left(\sum_{r=1}^n \int_\Omega |f| L_r \, \mathrm{d}\mu \right) \left\| t' - t \right\|.$$

$\square$

REMARK 4.2. *In Theorem 4.2, we may relax the requirements to hold for all $x \in \Omega$ μ-almost everywhere only.*

From Theorem 4.2, we can derive the following well-known result, which is useful for gradient-based optimization methods.

COROLLARY 4.2.1. *Let $(f, \Omega, \mathcal{A}, (\mathrm{Pr}_\theta)_{\theta \in \Theta})$ be a relaxation as defined in Definition 2.1, with the properties*

- *Θ is an open subset of $\mathbb{R}^n$,*
- *the measures have θ-Lipschitz differentiable densities with respect some measure μ, i.e. $\mathrm{Pr}_\theta = k_\theta \mu$ for all $\theta \in \Theta$, and*
- *the regularity assumptions of Theorem 4.2 with $\gamma \equiv \mathrm{id}_\Theta$ are fulfilled.[3]*

Then, $\nabla_\theta \mathbb{E}.(f) \equiv \mathbb{E}_\mu(f \nabla_\theta k.) \equiv \mathbb{E}.(f \nabla_\theta \ln k.)$ is Lipschitz.

---

[3]id_Θ denotes the identity map of Θ.

PROOF. This result is an application of the "log-likelihood trick".

$$\nabla_\theta \mathbb{E}.(f) \equiv \int_\Omega f(x) \nabla_\theta k.(x) \, \mu(\mathrm{d}x) \qquad \text{(by Theorem 4.2)}$$

$$\equiv \int_{\text{supp} \, k.} f(x)(\nabla_\theta k.)(x) \frac{k.(x)}{k.(x)} \, \mu(\mathrm{d}x) \qquad (\tfrac{k.(x)}{k.(x)} \equiv 1)$$

$$\equiv \int_{\text{supp} \, k.} f(x)(\nabla_\theta \ln k.)(x) k.(x) \, \mu(\mathrm{d}x)$$

$$\text{(log-likelihood trick)}$$

$$\equiv \int_{\text{supp} \, k.} f(x)(\nabla_\theta \ln k.)(x) \, \mathrm{Pr}.(\mathrm{d}x)$$

$$\text{(by definition } k.\mu \equiv \mathrm{Pr}.)$$

$$\equiv \mathbb{E}.(f \nabla_\theta \ln k.)$$

$$\text{(definition expected value, } \ln 0 := 0 \text{ on null sets)}$$

$\square$

Especially in black-box, noisy, or unstructured problems f the optimization of $\mathbb{E}.(f)$ using gradient-based methods is a reasonable alternative to finite-difference methods. The transfer of Lipschitzness of gradients can be exploited for improved convergence guarantees [2]. In the literature, the properties derived in this section have been observed experimentally for many algorithms employing stochastic relaxations at their core, specifically in the setting of Example 2.1.

5 WEAK CONVEXITY

The notion of convexity of a function is central to continuous optimization. Non-convex optimization problems are much harder to solve in general. However, for families of translated probability measures—including multivariate normal distributions—we show that a function's convexity is preserved in its relaxation. More importantly, we provide conditions under which non-convex functions result in convex stochastic relaxations. To this end, consider the following class of functions f in case of $\Omega = \Theta = \mathbb{R}^n$.

DEFINITION 5.1 (WEAK CONVEXITY). *Let a relaxation as defined in Definition 2.1 be given by $(f, \mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathrm{Pr}_\theta)_{\theta \in \mathbb{R}^n})$ with*

$$\mathrm{Pr}_\theta(A) = \mathrm{Pr}_0(A - \theta) \text{ for all } \theta \in \mathbb{R}^n \text{ and for all } A \in \mathcal{B}(\mathbb{R}^n).$$

*We call f (strictly/ m-strongly) Pr_0-**weakly convex** if there is $\sigma^* > 0$ such that for all $\sigma > \sigma^*$*

$$\mathrm{Pr}_\theta{}^\sigma(A) := \mathrm{Pr}_0(\sigma^{-1}(A - \theta)) \text{ for all } \theta \in \mathbb{R}^n \text{ and for all } A \in \mathcal{B}(\mathbb{R}^n)$$

gives a (strictly/ m-strongly) convex relaxation $\theta \in \mathbb{R}^n \mapsto \mathbb{E}_{\theta,\sigma}(f)$.

As a direct result, we obtain the following by substitution.

LEMMA 5.1. *In the setting of Definition 5.1, if the probability measure Pr_0 has a density k_0 with respect to the Lebesgue measure λ, then for all $\sigma > 0$, for all $\theta \in \mathbb{R}^n$, and for all $A \in \mathcal{B}(\mathbb{R}^n)$, we have*

$$\mathrm{Pr}_\theta{}^\sigma(A) = \mathrm{Pr}_0(\sigma^{-1}(A - \theta)) \qquad \text{(definition } \mathrm{Pr}_\theta{}^\sigma)$$

$$= \int_{\sigma^{-1}(A-\theta)} k_0 \, \mathrm{d}\lambda \qquad \text{(definition density)}$$

$$= \int_A \sigma^{-n} k_0(\sigma^{-1}(\cdot - \theta)) \, \mathrm{d}\lambda. \qquad \text{(substitution)}$$

That is, $k_{\theta,\sigma} := \sigma^{-n} k_0(\sigma^{-1}(\cdot - \theta))$ is a Lebesgue density of $\mathrm{Pr}_\theta{}^\sigma$.

The goal for the rest of the section will be the derivation of a weak convexity property for stochastic relaxations. The approach we take will be based on Fourier analysis and will work with a relaxation of the superposition $f = r + g : \mathbb{R}^n \to \mathbb{R}$ based on a class of densities that are rapidly decreasing. The function r will be assumed to be strongly convex, and g can be understood as a disturbance. While in Example 2.1, g was a finite sum of cosines and the densities were isotropic Gaussian, we aim to find a general setting that provides interpretation for the success of stochastic relaxations in global and robust optimization.

THEOREM 5.1 (PRESERVING CONVEXITY). *Let a relaxation as defined in Definition 2.1 be given by $(f, \mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\Pr_\theta)_{\theta \in \mathbb{R}^n})$ with*

- $\Pr_\theta(A) = \Pr_0(A - \theta)$ *for all $\theta \in \mathbb{R}^n$ and for all $A \in \mathcal{B}(\mathbb{R}^n)$, and*
- *f is strongly convex, i.e., there exists $m > 0$ such that for all $v, w \in \mathbb{R}^n$ and for all $t \in [0, 1]$*

$$f(tv + (1-t)w) \leq tf(v) + (1-t)f(w) - \tfrac{1}{2}mt(1-t)\|v - w\|^2 .$$

Then, the relaxation $\theta \in \mathbb{R}^n \mapsto \mathbb{E}_\theta(f)$ is m-strongly convex.

PROOF. Let $\psi(\theta, x) := \psi_\theta(x) := x + \theta$ for all $x \in \mathbb{R}^n$ and for all $\theta \in \mathbb{R}^n$. Further let $v, w \in \mathbb{R}^n$ and $t \in [0, 1]$, then

$$\mathbb{E}_{tv+(1-t)w}(f) = \int_{\mathbb{R}^n} f \, d\Pr_{tv+(1-t)w} \quad \text{(definition expected value)}$$

$$= \int_0^\infty \Pr_{tv+(1-t)w}(\{x \in \mathbb{R}^n \mid f(x) > t\}) \, dt$$
$$\text{(definition Lebesgue integral)}$$

$$= \int_0^\infty \Pr_0\big(\psi_{tv+(1-t)w}^{-1}(\{x \in \mathbb{R}^n \mid f(x) > t\})\big) \, dt$$
$$\text{(assumption)}$$

$$= \int_0^\infty \Pr_0(\{x \in \mathbb{R}^n \mid (f \circ \psi_{tv+(1-t)w})(x) > t\}) \, dt$$
$$\text{(definition inverse)}$$

$$= \int_{\mathbb{R}^n} f(x + tv + (1-t)w) \Pr_0(dx)$$
$$\text{(definition Lebesgue integral)}$$

$$= \int_{\mathbb{R}^n} f(t(x+v) + (1-t)(x+w)) \Pr_0(dx)$$
$$(x = tx + (1-t)x)$$

$$\leq \int_{\mathbb{R}^n} tf(x+v) + (1-t)f(x+w)$$
$$- \tfrac{1}{2}mt(1-t)\|v-w\|^2 \Pr_0(dx)$$
$$\text{(third assumption)}$$

$$= t \int_{\mathbb{R}^n} f(x+v) \Pr_0(dx)$$

$$+ (1-t) \int_{\mathbb{R}^n} f(x+w) \Pr_0(dx)$$
$$- \tfrac{1}{2}mt(1-t)\|v-w\|^2$$
$$\text{(linearity, probability measure)}$$

$$= t\,\mathbb{E}_v(f) + (1-t)\,\mathbb{E}_w(f) - \tfrac{1}{2}mt(1-t)\|v-w\|^2 .$$
$$\text{(first steps reversed)}$$
□

REMARK 5.1. *Preservation also holds for (strictly) convex functions.*

The following result is developed in preparation of Theorem 5.2. To retain the generality of this section and obtain a convexity result similar to Example 2.1, the distributional formulation of the Fourier transform is needed. This result will pave the way for very practical corollaries at the end of this section.

LEMMA 5.2. *Let λ be the Lebesgue measure*

- *$g : \mathbb{R}^n \to \mathbb{R}$ measurable with $T_g \in \mathcal{S}'(\mathbb{R}^n)$,[4]*
- *$\mathcal{F}(T_g)(\varphi) = 0$ for all $\varphi \in C_c^\infty(U_\varepsilon(0))$ for some $\varepsilon > 0$, and*

Given $h \in \mathcal{S}(\mathbb{R}^n)$ and $\sigma \in \mathbb{R}$, then for all $y \in \mathbb{R}^n$ uniformly

$$T_g\big(\sigma^{-n}h(\sigma^{-1}(\cdot - y))\big) = \int_{\mathbb{R}^n} g(x)\sigma^{-n}h(\sigma^{-1}(x-y))\,\lambda(dx) \xrightarrow{\sigma\to\infty} 0 .$$

PROOF. Applying general properties of the Fourier transform, we get the identity

$$T_g\big(\sigma^{-n}h(\sigma^{-1}(\cdot - y))\big) = (\mathcal{F}^{-1}(\mathcal{F}T_g))\big(\sigma^{-n}h(\sigma^{-1}(\cdot - y))\big)$$
$$\text{(Fourier inversion for distributions)}$$

$$= (\mathcal{F}T_g)\big(\mathcal{F}^{-1}\sigma^{-n}h(\sigma^{-1}(\cdot - y))\big)$$
$$\text{(definition inverse transform)}$$

$$= (\mathcal{F}T_g)\big(e^{i\langle y,\cdot\rangle}\sigma^{-n}\mathcal{F}^{-1}(h(\sigma^{-1}\cdot))\big)$$
$$\text{(shift and linearity)}$$

$$= (\mathcal{F}T_g)\big(e^{i\langle y,\cdot\rangle}\mathcal{F}^{-1}(h)(\sigma\cdot)\big) . \quad \text{(scaling)}$$

The Fourier transform is an automorphism on Schwartz space, therefore $s := \mathcal{F}^{-1}(h)$ is a Schwartz function. Now, due to continuity of $\mathcal{F}T_g$, we would like to have

$$e^{i\langle y,\cdot\rangle}s(\sigma\cdot) \xrightarrow{\sigma\to\infty} 0 \text{ for all } y \in \mathbb{R}^n \text{ uniformly in } \mathcal{S}.$$

This, however, does not hold in general. We need the assumption $\mathcal{F}T_{g|U_\varepsilon(0)} \equiv 0$.
To this end, consider a smooth partition of unity $u_1, u_2 : \mathbb{R}^n \to \mathbb{R}$ with $u_1 + u_2 \equiv 1$ such that $\operatorname{supp} u_1 \subseteq \overline{U_\varepsilon(0)}$ and $\operatorname{supp} u_2 \subseteq \mathbb{R}^n \setminus U_{\varepsilon/2}(0)$ and see that

$$\mathcal{F}T_g\big(e^{i\langle y,\cdot\rangle}s(\sigma\cdot)\big) \equiv \mathcal{F}T_g\big((u_1 + u_2)e^{i\langle y,\cdot\rangle}s(\sigma\cdot)\big) \quad \text{(definition } u_1, u_2)$$

$$\equiv \mathcal{F}T_g\big(u_1 e^{i\langle y,\cdot\rangle}s(\sigma\cdot)\big) + \mathcal{F}T_g\big(u_2 e^{i\langle y,\cdot\rangle}s(\sigma\cdot)\big)$$
$$\text{(linearity)}$$

$$\equiv 0 + \mathcal{F}T_g\big(u_2 e^{i\langle y,\cdot\rangle}s(\sigma\cdot)\big) . \quad \text{(assumption)}$$

Due to s being a Schwartz function and $\big|e^{i\langle y,\cdot\rangle}\big| \equiv 1$ for all $y \in \mathbb{R}^n$, we have

$$u_2 e^{i\langle y,\cdot\rangle}s(\sigma\cdot) \xrightarrow{\sigma\to\infty} 0 \text{ for all } y \in \mathbb{R}^n \text{ uniformly in } \mathcal{S}.$$

By continuity of $\mathcal{F}T_g$ this implies the statement of the theorem. □

Now, we will describe a setting for which the relaxed problem is convex when the original cost function may not be. As described previously, we consider the case where the function f is a strongly convex function superimposed with a disturbance, impairing convexity and thus global optimization.

---

[4] T_g denotes the functional $\varphi \mapsto \int_{\mathbb{R}^n} g\varphi \, d\lambda$. This property ensures that the integral exists. Continuity and linearity is generally fulfilled. $\mathcal{S}$ denotes the Schwartz space and $\mathcal{S}'$ its dual space.

THEOREM 5.2 (FILTERING). *Let λ be the Lebesgue measure and $f : \mathbb{R}^n \to \mathbb{R}$ be a function that admits a decomposition*

$$f \equiv r + g$$

such that

- *r is m-strongly convex,*
- *$\mathcal{F}(g)_{|U_\varepsilon(0)} \equiv 0$ in the distributional sense for some $\varepsilon > 0$, and*
- *$rs, gs \in \mathcal{L}^1(\mathbb{R}^n)$ for all Schwartz functions $s \in \mathcal{S}(\mathbb{R}^n)$.*

Then for every non-negative, somewhere non-zero Schwartz function $k \in \mathcal{S}(\mathbb{R}^n)$ with $c_k := \left(\int_{\mathbb{R}^n} k \, d\lambda \right)^{-1}$ and $m^* < m$, f is m^*-strongly $(c_k k)\lambda$-weakly convex, i.e., there exists σ_k^* such that for all $\sigma > \sigma_k^*$

$$\theta \in \mathbb{R}^n \longmapsto \mathbb{E}_{\theta,\sigma}(f) = \int_{\mathbb{R}^n} f(x) c_k \sigma^{-n} k(\sigma^{-1}(x - \theta)) \, \lambda(dx)$$

is m^*-strongly convex.

PROOF. Pick a non-negative, somewhere non-zero Schwartz function $k \in \mathcal{S}(\mathbb{R}^n)$. For all $\sigma > 0$ and for all $\theta \in \mathbb{R}^n$, we have

$$\mathbb{E}_{\theta,\sigma}(f) = \mathbb{E}_{\theta,\sigma}(r) + \mathbb{E}_{\theta,\sigma}(g)$$

and we know by direct application of Theorem 5.1 that $\theta \in \mathbb{R}^n \mapsto \mathbb{E}_{\theta,\sigma}(r)$ is strongly convex with the same constant as r for all $\sigma > 0$. We prove the result by showing that for all $\varepsilon > 0$ there exists σ_ε such that all partial second derivatives of $\mathbb{E}_{\theta,\sigma_\varepsilon}(g)$ are uniformly bounded by ε. To see why this is, observe the following arguments.

- For all $i, j \in \mathbb{N}_{\leq n}$, $\sigma > 0$, $\theta \in \mathbb{R}^n$, and $x \in \mathbb{R}^n$ there is the identity

$$\partial_{\theta_i, \theta_j} k(\sigma^{-1}(x - \theta)) = \sigma^{-2}(\partial_{ij} k)(\sigma^{-1}(x - \theta)) .$$

- By Theorem 4.2, we have

$$\partial_{\theta_i, \theta_j} \mathbb{E}_{\theta,\sigma}(g) = \sigma^{-2} \int_{\mathbb{R}^d} g(x) \sigma^{-n} c_k (\partial_{ij} k)(\sigma^{-1}(x - \theta)) \, \lambda(dx) .$$

- The function $x \in \mathbb{R}^n \mapsto c_k(\partial_{ij} k)(x)$ is again a Schwartz function and together with g fulfills the assumptions of Lemma 5.2 for all $i, j \in \mathbb{N}_{\leq n}$. This provides a uniform bound of ε for all $\sigma > \sigma_{\varepsilon,i,j}$ on

$$\int_{\mathbb{R}^d} g(x) \sigma^{-n} c_k (\partial_{ij} k)(\sigma^{-1}(x - \theta)) \, \lambda(dx)$$

and thereby due to the factor σ^{-2} and without loss of generality on $\partial_{\theta_i, \theta_j} \mathbb{E}_{\theta,\sigma}(g)$.

- By picking σ_ε^*, the largest of $(\sigma_{\varepsilon,i,j})_{i,j \in \mathbb{N}_{\leq n}}$, one obtains a uniform bound of ε on all second partial derivatives of $\theta \in \mathbb{R}^n \mapsto \mathbb{E}_{\theta,\sigma_\varepsilon^*}(g)$.

It remains to be shown that if σ is picked large enough, this implies strong convexity for an arbitrary parameter m^* smaller than the convexity parameter of r, namely m. For all $x \in \mathbb{R}^n$ we want to have

$$x^T (H(\mathbb{E}_{\cdot,\sigma}(r)) - m^* I_n + \varepsilon 1_n) x \geq 0 ,$$

where $H(\mathbb{E}_{\cdot,\sigma}(r))$ denotes the Hessian of $\mathbb{E}_{\cdot,\sigma}(r)$, I_n is the identity, and 1_n is a matrix such that

$$|1_{n,ij}| \leq 1 \text{ for all } i, j.$$

$\varepsilon 1_n$ corresponds to the Hessian of $\mathbb{E}_{\cdot,\sigma}(g)$. We have

$$x^T (H(\mathbb{E}_{\cdot,\sigma}(r)) - m I_n) x \geq 0$$

by Theorem 5.1 and, therefore, we show

$$x^T ((m - m^*) I_n + \varepsilon 1_n) x \geq (m - m^*) \max_i |x_i|^2 - \varepsilon n^2 \max_i |x_i|^2$$

$$= \max_i |x_i|^2 (m - m^* - \varepsilon n^2) \geq 0 .$$

Which is true for $\varepsilon \leq \frac{m - m^*}{n^2}$. □

Theorem 5.2 suggests non-trivial sufficient conditions for functions on the reals whose stochastic relaxations turn out to be convex functions when paired with probability measures that have Schwartz densities. There are a number of handy results we obtain with Theorem 5.2.

COROLLARY 5.2.1 (DETERMINISTIC COSINE DISTURBANCE). *Let $f = r + g : \mathbb{R}^n \to \mathbb{R}$ be the sum of a polynomially bounded, m-strongly convex function r and for all $x \in \mathbb{R}^n$*

$$g(x) = \sum_{j=1}^{\infty} a_j \cos(\langle \xi_j, x \rangle + \psi_j)$$

with

- *$a_j \in \mathbb{R}$ with $\sum_j |a_j| < \infty$, and*
- *$\xi_j \in \mathbb{R}^n$ such that there exists $\varepsilon > 0$ with $\|\xi_j\| \geq \varepsilon$ for all $j \in \mathbb{N}$.*

Let Pr_k be a probability measure with a rapidly decreasing Lebesgue density k and $m^* < m$. Then, f is m^*-strongly Pr_k-weakly convex.

PROOF. We prove that g fulfills $\mathcal{F}(g)_{|U_\varepsilon(0)} \equiv 0$ in the distributional sense and apply Theorem 5.2. We have for all $j \in \mathbb{N}$

$$\cos(\langle \xi_j, x \rangle + \psi_j) = \frac{e^{i(\langle \xi_j, x \rangle + \psi_j)} + e^{-i(\langle \xi_j, x \rangle + \psi_j)}}{2}$$

$$= \frac{e^{i\psi_j} e^{i\langle \xi_j, x \rangle} + e^{-i\psi_j} e^{i\langle -\xi_j, x \rangle}}{2} . \quad (5)$$

Therefore, for all Schwartz functions $\varphi \in \mathcal{S}(\mathbb{R}^n)$, we get

$$\mathcal{F} T_g(\varphi) = T_g(\mathcal{F}\varphi) = \int_{\mathbb{R}^n} g \mathcal{F}\varphi \, d\lambda$$

(definition distributional Fourier transform)

$$= \sum_{j=1}^{\infty} a_j \int_{\mathbb{R}^n} \cos(\langle \xi_j, x \rangle + \psi_j) \mathcal{F}\varphi \, \lambda(dx)$$

(definition g, linearity)

$$= \sum_{j=1}^{\infty} \frac{a_j}{2} \left(e^{i\psi_j} \int_{\mathbb{R}^n} e^{i\langle \xi_j, x \rangle} \mathcal{F}\varphi \, \lambda(dx) \right.$$

$$\left. + e^{-i\psi_j} \int_{\mathbb{R}^n} e^{i\langle -\xi_j, x \rangle} \mathcal{F}\varphi \, \lambda(dx) \right)$$

(Equation (5), linearity)

$$= \sum_{j=1}^{\infty} \frac{a_j}{2} \left(e^{i\psi_j} \mathcal{F}^{-1}(\mathcal{F}\varphi)(\xi_j) + e^{-i\psi_j} \mathcal{F}^{-1}(\mathcal{F}\varphi)(-\xi_j) \right)$$

(definition inverse Fourier transform)

$$= \sum_{j=1}^{\infty} \frac{a_j}{2} \left(e^{i\psi_j} \varphi(\xi_j) + e^{-i\psi_j} \varphi(-\xi_j) \right)$$

(Fourier inversion theorem)

This implies $\mathcal{F}T_g = \sum_{j=1}^{\infty} \frac{a_j}{2}\left(e^{i\psi_j}\delta_{\xi_j} + e^{-i\psi_j}\delta_{-\xi_j}\right)$, where δ is the Dirac distribution. The assumption $\|\xi_j\| > \varepsilon$ for all $j \in \mathbb{N}$ therefore implies $\mathcal{F}(g)_{|U_\varepsilon(0)} \equiv 0$. □

Corollary 5.2.1 also recovers the convexity result from Example 2.1.

COROLLARY 5.2.2. *The Rastrigin function*

$$x \in \mathbb{R}^n \longmapsto f(x) := \|x\|^2 - \sum_{j=1}^n a_j \cos(\xi_j x_j),$$

where $\xi_j, a_j > 0$ for all $j \in \mathbb{N}_{\le n}$, is m^*-strongly $\mathcal{N}(0,1)$-weakly convex for all $m^* < 1$.

We also obtain the following result for stochastic disturbances. From the algorithmic perspective, the following model constitutes static noise.

COROLLARY 5.2.3 (STOCHASTIC COSINE DISTURBANCE). *Let $f = r + g : \mathbb{R}^n \to \mathbb{R}$ be the sum of a polynomially bounded, m-strongly convex function r and a random field*

$$g(x) = \sum_{j=1}^{\infty} A_j \cos(\langle \xi_j, x\rangle + \psi_j)$$

with

- *$(A_j)_{j \in \mathbb{N}}$ are real-valued random variables on the same probability space such that $\sum_j |A_j| < C$ with probability $p \in [0,1]$, and*
- *$\xi_j \in \mathbb{R}^n$ such that there exists $\varepsilon > 0$ with $\|\xi_j\| \ge \varepsilon$ for all $j \in \mathbb{N}$.*

Let $\Pr_k$ be a probability measure with a rapidly decreasing Lebesgue density k and $m^* < m$. Then, f is m^*-strongly $\Pr_k$-weakly convex with at least probability p with the same threshold scaling constant σ^*. Further, $\sigma^*(C) \in O\left(\sqrt[d]{C}\right)$ for all $d \in \mathbb{N}$.

PROOF. Analogously to Theorem 5.2 and Lemma 5.2 the result follows in conjunction with the following bound. For all Schwartz functions $\varphi \in \mathcal{S}(\mathbb{R}^n)$, we have with probability p

$$\mathcal{F}T_g(\varphi) = \sum_{j=1}^{\infty} \frac{A_j}{2}\left(e^{i\psi_j}\varphi(\xi_j) + e^{-i\psi_j}\varphi(-\xi_j)\right)$$

(proof of Corollary 5.2.1)

$$\implies \left|\mathcal{F}T_g(\varphi)\right| \le \sum_{j=1}^{\infty} \frac{|A_j|}{2}\left(\left|\varphi(\xi_j)\right| + \left|\varphi(-\xi_j)\right|\right) \quad \text{(triangle inequality)}$$

$$\le \sum_{j=1}^{\infty} |A_j| \sup_{x \in U_\varepsilon(0)^C} |\varphi(x)| \quad (\varphi \text{ is Schwartz})$$

$$\le C \sup_{x \in U_\varepsilon(0)^C} |\varphi(x)|. \quad (\text{assumption } \sum_j |A_j| < C)$$

By the fact that φ is rapidly decreasing (see Lemma 5.2 for why φ can be considered fixed), we also have

$$C \sup_{x \in U_\varepsilon(0)^C} |\varphi(\sigma x)| \le C\frac{B}{\sigma^d}$$

for some $B \ge 0$ and for all $\sigma > 0$ and for all $d \in \mathbb{N}$. Therefore, as we want to pick σ^* to bound the value of $\left|\mathcal{F}T_g(\sigma^*\varphi)\right|$ by say some $c > 0$ dependent on the convexity parameter of r, in the spirit of

Theorem 5.2 and Lemma 5.2, we have $\sigma^*(C) \in O\left(\sqrt[d]{C}\right)$ for all $d \in \mathbb{N}$ due to

$$C\frac{B}{\sigma^d} = c \iff \sigma = \sqrt[d]{B/c}\sqrt[d]{C}.$$

□

6 CONCLUSION

We established that stochastic relaxations have many favorable properties for optimization in surprisingly general settings. Concretely, we provided mild conditions under which optima are persistent under stochastic relaxation and thereby established the approach's consistency, which aligns with our intuition. Next to consistency, we have also shown that probability measures with Lipschitz differentiable densities under mild regularity conditions result in relaxations with Lipschitz differentials, allowing favorable convergence properties for first-order optimization methods. However, the arguably most important contributions are the convexity conditions for stochastic relaxations derived in the last section, which offer the possibility of principled global optimization under high-frequency, even stochastic, disturbances. Beyond that, we hope that the insight developed will be useful in furthering state-of-the-art optimization methods.

Future Work. We believe that some of the most promising directions for future work lie in extending the convexity results for stochastic relaxations based on other methods from harmonic analysis. In practice, particularly in evolution strategies, the decoupling of the relaxation and (numerical) integration methods suggested by the setting at hand could offer the opportunity to design many new optimization methods. The gradient representation of relaxations of differentiable functions likely has powerful applications in problems where gradients are available, such as in gray-box problems in machine learning. Whether convexity results for relaxations hold for probability measures with a support on Lebesgue null sets may also pose an interesting question. These measures can offer strongly reduced computational cost in numerical integration—related algorithms have recently yielded spectacular results [3, 5]. Connecting the convexity results with existing approaches to control the scaling of probability measures during an optimization process is also an important next step.

REFERENCES

[1] Hans-Georg Beyer. 2014. Convergence Analysis of Evolutionary Algorithms That Are Based on the Paradigm of Information Geometry. *Evolutionary Computation* 22, 4 (2014), 679–709.

[2] Sébastien Bubeck. 2015. Convex Optimization: Algorithms and Complexity. *Found. Trends Mach. Learn.* 8, 3–4 (2015), 231–357.

[3] Krzysztof Choromanski, Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, and Vikas Sindhwani. 2019. From Complexity to Simplicity: Adaptive ES-Active Subspaces for Blackbox Optimization. *Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019)* (2019).

[4] Marco Loog, Johannes Jisse Duistermaat, and Luc M. J. Florack. 2001. On the Behavior of Spatial Critical Points under Gaussian Blurring A Folklore Theorem and Scale-Space Constraints. In *Scale-Space and Morphology in Computer Vision.* 183–192.

[5] I. Loshchilov, T. Glasmachers, and H. Beyer. 2019. Large Scale Black-Box Optimization by Limited-Memory Matrix Adaptation. *IEEE Transactions on Evolutionary Computation* 23, 2 (2019), 353–358.

[6] Alvaro Maggiar, Andreas Wächter, Irina S. Dolinskaya, and Jeremy Staum. 2018. A Derivative-Free Trust-Region Algorithm for the Optimization of Functions Smoothed via Gaussian Convolution Using Adaptive Multiple Importance Sampling. *SIAM Journal on Optimization* 28, 2 (2018), 1478–1507.

[7] Luigi Malago, Matteo Matteucci, and Pistone Giovanni. 2009. Stochastic Relaxation as a Unifying Approach in 0/1 Programming. In *NIPS 2009 Workshop on*

Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML 2009). 1–7.

[8] Hossein Mobahi and John W Fisher III. 2015. On the link between gaussian homotopy continuation and convex envelopes. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition.* 43–56.

[9] Hossein Mobahi and John W Fisher III. 2015. A theoretical analysis of optimization by gaussian continuation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.* 1205–1211.

[10] Hossein Mobahi and Yi Ma. 2012. Gaussian smoothing and asymptotic convexity. *Coordinated Science Laboratory Report no. UILU-ENG-12-2201, DC-254* (2012).

[11] Yurii Nesterov and Vladimir Spokoiny. 2017. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics* 17, 2 (2017), 527–566.

[12] Yann Ollivier, Ludovic Arnold, Anne Auger, and Nikolaus Hansen. 2017. Information-geometric optimization algorithms: A unifying picture via invariance principles. *The Journal of Machine Learning Research* 18, 1 (2017), 564–628.

[13] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. 2017. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864* (2017).

[14] René L. Schilling. 2005. *Measures, Integrals and Martingales.* Cambridge University Press.

[15] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber. 2008. Natural Evolution Strategies. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence).* 3381–3387.

[16] Jingzhao Zhang, Hongyi Zhang, and Suvrit Sra. 2018. R-SPIDER: A Fast Riemannian Stochastic Optimization Algorithm with Curvature Independent Rate. *arXiv preprint arXiv:1811.04194* (2018).