

## A PROOFS

### A.1 PROOF OF LEMMA 3

*Proof of Lemma 3.* Let  $\Omega \subset \mathbb{R}^n$  be the set of values where  $\nabla_\epsilon \mu(\epsilon)$  is undefined.  $\mu$  is differentiable a.e. and  $\Omega$  has Lebesgue measure 0.

Recapitulating (5) from the proof of Lemma 1, we have

$$\nabla_x f_\epsilon(x) = - \int f(x + \epsilon) \nabla_\epsilon \mu(\epsilon) d\epsilon. \quad (14)$$

Replacing  $\nabla_\epsilon \mu(\epsilon)$  by *any* of the weak derivatives  $\nu$  of  $\mu$ , which exists and is integrable due to absolute continuity, we have

$$\nabla_x f_\epsilon(x) = - \int f(x + \epsilon) \nu(\epsilon) d\epsilon \quad (15)$$

$$= - \int_{\mathbb{R}^n \setminus \Omega} f(x + \epsilon) \nu(\epsilon) d\epsilon - \int_{\Omega} f(x + \epsilon) \nu(\epsilon) d\epsilon. \quad (16)$$

Because  $\mu$  is absolutely continuous and as the Lebesgue measure of  $\Omega$  is 0, per Hölder’s inequality

$$\int_{\Omega} |f(x + \epsilon) \nu(\epsilon)| d\epsilon \leq \int_{\Omega} |f(x + \epsilon)| d\epsilon \cdot \int_{\Omega} |\nu(\epsilon)| d\epsilon = \int_{\Omega} |f(x + \epsilon)| d\epsilon \cdot 0 = 0 \quad (17)$$

where  $\int_{\Omega} |\nu(\epsilon)| d\epsilon = 0$  follows from absolute continuity of  $\mu$ . Thus,

$$\int_{\Omega} f(x + \epsilon) \nu(\epsilon) d\epsilon = 0. \quad (18)$$

As  $\nu = \nabla_\epsilon \mu(\epsilon)$  for all  $\epsilon \in \mathbb{R}^n \setminus \Omega$

$$\nabla_x f_\epsilon(x) = - \int_{\mathbb{R}^n \setminus \Omega} f(x + \epsilon) \nu(\epsilon) d\epsilon - \int_{\Omega} f(x + \epsilon) \nu(\epsilon) d\epsilon = - \int_{\mathbb{R}^n \setminus \Omega} f(x + \epsilon) \nabla_\epsilon \mu(\epsilon) d\epsilon, \quad (19)$$

showing that for all possible choices of  $\nu$ , the gradient estimator coincides. Thus, we complete our proof via

$$\nabla_x f_\epsilon(x) = - \int_{\mathbb{R}^n \setminus \Omega} f(x + \epsilon) \mu(\epsilon) \nabla_\epsilon \log \mu(\epsilon) d\epsilon = \mathbb{E}_{\epsilon \sim \mu} \left[ f(x + \epsilon) \cdot \mathbf{1}_{\epsilon \notin \Omega} \cdot \nabla_\epsilon \log \mu(\epsilon) \right]. \quad (20)$$

After completing the proof, we remark that, if the density was not continuous, e.g., uniform  $\mathcal{U}([0, 1])$ , then  $\int_{\{0\}} \nabla_\epsilon \mu(\epsilon) d\epsilon = \left[ \mu(\epsilon) \right]_{\epsilon \nearrow 0}^{\epsilon \searrow 0} = 1$ . This means that the weak derivative is not defined (or loosely speaking “the derivative is infinity”), thereby violating the assumptions of Hölder’s inequality (Eq. 17). This concludes that continuity is required for the proof to hold.  $\square$

## A.2 PROOF OF LEMMA 6

*Proof of Lemma 6.*

$$\nabla_\gamma f_{\gamma\epsilon}(x) = \nabla_\gamma \mathbb{E}_{\epsilon \sim \mu} [f(x + \gamma \cdot \epsilon)] \quad (21)$$

$$= \nabla_\gamma \int f(x + \gamma \cdot \epsilon) \mu(\epsilon) d\epsilon \quad (22)$$

$$(u = x + \epsilon \cdot \gamma \Rightarrow \epsilon = \frac{u-x}{\gamma} ; \frac{du}{d\epsilon} = \gamma \Rightarrow d\epsilon = \frac{1}{\gamma} du) \quad (23)$$

$$= \nabla_\gamma \int f(u) \mu(\epsilon) \frac{1}{\gamma} du \quad (24)$$

$$= \int f(u) \nabla_\gamma (\mu(\epsilon) \frac{1}{\gamma}) du \quad (25)$$

$$= \int f(u) (\frac{1}{\gamma} \nabla_\gamma \mu(\epsilon) + \mu(\epsilon) \nabla_\gamma \frac{1}{\gamma}) du \quad (26)$$

$$= \int f(u) (\frac{1}{\gamma} (\nabla_\epsilon \mu(\epsilon))^\top \frac{\partial \epsilon}{\partial \gamma} - \mu(\epsilon) \frac{1}{\gamma^2}) du \quad (27)$$

$$= \int f(u) (\frac{1}{\gamma} (\nabla_\epsilon \mu(\epsilon))^\top \frac{\partial}{\partial \gamma} \frac{u-x}{\gamma} - \mu(\epsilon) \frac{1}{\gamma^2}) du \quad (28)$$

$$= \int f(u) (\frac{1}{\gamma} (\nabla_\epsilon \mu(\epsilon))^\top (-\frac{\epsilon}{\gamma}) - \frac{1}{\gamma^2} \mu(\epsilon)) du \quad (29)$$

$$= \int f(u) (-(\nabla_\epsilon \mu(\epsilon))^\top \epsilon - \mu(\epsilon)) \frac{1}{\gamma^2} du \quad (30)$$

$$\left( \nabla_\epsilon \log \mu(\epsilon) = \frac{1}{\mu(\epsilon)} \nabla_\epsilon \mu(\epsilon) \Rightarrow \nabla_\epsilon \mu(\epsilon) = \mu(\epsilon) \nabla_\epsilon \log \mu(\epsilon) \right) \quad (31)$$

$$= \int f(u) (-(\mu(\epsilon) \nabla_\epsilon \log \mu(\epsilon))^\top \epsilon - \mu(\epsilon)) \frac{1}{\gamma^2} \cdot \underbrace{\gamma d\epsilon}_{=du} \quad (32)$$

$$= \int f(u) \cdot (-(\nabla_\epsilon \log \mu(\epsilon))^\top \epsilon - 1) \cdot \frac{1}{\gamma} \cdot \mu(\epsilon) d\epsilon \quad (33)$$

$$= \int f(u) \cdot (-1 + (\nabla_\epsilon - \log \mu(\epsilon))^\top \epsilon) \cdot \frac{1}{\gamma} \cdot \mu(\epsilon) d\epsilon \quad (34)$$

$$= \mathbb{E}_{\epsilon \sim \mu} [f(x + \gamma \cdot \epsilon) \cdot (-1 + (\nabla_\epsilon - \log \mu(\epsilon))^\top \epsilon) / \gamma] . \quad (35)$$

□

## A.3 PROOF OF THEOREM 7

*Proof of Theorem 7.*

**Part 1:**  $\partial f_{\mathbf{L}\epsilon}(x) / \partial x$

We perform a change of variables,  $u = x + \mathbf{L}\epsilon \implies \epsilon = \mathbf{L}^{-1}(u - x)$  and

$$d\epsilon = \frac{du}{d\mathbf{L}} d\mathbf{L} = \frac{d\epsilon}{du} du = \frac{d\mathbf{L}^{-1}(u - x)}{du} du = \frac{d\mathbf{L}^{-1}u}{du} du = \det(\mathbf{L}^{-1}) du \quad (36)$$

Thus,

$$f_{\mathbf{L}\epsilon}(x) = \int f(x + \mathbf{L}\epsilon) \mu(\epsilon) d\epsilon = \int f(u) \cdot \mu(\mathbf{L}^{-1}(u - x)) \cdot \det(\mathbf{L}^{-1}) du . \quad (37)$$

Now,

$$\nabla_x f_{\mathbf{L}\epsilon}(x)_i = \nabla_x \int f(u)_i \cdot \mu(\mathbf{L}^{-1}(u - x)) \cdot \det(\mathbf{L}^{-1}) du \quad (38)$$

$$= \int f(u)_i \cdot \nabla_x (\mu(\mathbf{L}^{-1}(u - x))) \cdot \det(\mathbf{L}^{-1}) du \quad (39)$$

$$= \int f(u)_i \cdot \mathbf{L}^{-1} \cdot (\nabla_{\epsilon} \mu(\epsilon)) \cdot \det(\mathbf{L}^{-1}) du \quad (40)$$

$$= \int f(x + \mathbf{L}\epsilon)_i \cdot \mathbf{L}^{-1} \cdot \nabla_{\epsilon} \mu(\epsilon) d\epsilon \quad (41)$$

$$= \int f(x + \mathbf{L}\epsilon)_i \cdot \mathbf{L}^{-1} \cdot \mu(\epsilon) \cdot \nabla_{\epsilon} \log \mu(\epsilon) d\epsilon \quad (42)$$

$$= \mathbb{E}_{\epsilon \sim \mu} [f(x + \mathbf{L}\epsilon)_i \cdot \mathbf{L}^{-1} \cdot \nabla_{\epsilon} \log \mu(\epsilon)] \quad (43)$$

## Part 2: $\partial f_{\mathbf{L}\epsilon}(x) / \partial \mathbf{L}$

We use the same change of variables as above.

$$\nabla_{\mathbf{L}} \mathbb{E}_{\epsilon \sim \mu} [f(x + \mathbf{L} \cdot \epsilon)_i] \quad (44)$$

$$= \nabla_{\mathbf{L}} \int f(x + \mathbf{L}\epsilon)_i \cdot \mu(\epsilon) d\epsilon \quad (45)$$

$$= \nabla_{\mathbf{L}} \int f(u)_i \cdot \mu(\mathbf{L}^{-1}(u - x)) \cdot \det(\mathbf{L}^{-1}) du \quad (46)$$

$$= \int f(u)_i \cdot \nabla_{\mathbf{L}} (\mu(\mathbf{L}^{-1}(u - x)) \cdot \det(\mathbf{L}^{-1})) du \quad (47)$$

$$= \int f(x + \mathbf{L}\epsilon)_i \cdot \nabla_{\mathbf{L}} (\mu(\mathbf{L}^{-1}(u - x)) \cdot \det(\mathbf{L}^{-1})) / \det(\mathbf{L}^{-1}) d\epsilon \quad (48)$$

$$= \mathbb{E}_{\epsilon \sim \mu} [f(x + \mathbf{L}\epsilon)_i \cdot \nabla_{\mathbf{L}} (\mu(\mathbf{L}^{-1}(u - x)) \cdot \det(\mathbf{L}^{-1})) \cdot \det(\mathbf{L}) / \mu(\epsilon)] \quad (49)$$

Now, while  $\nabla_{\mathbf{L}} (\mu(\mathbf{L}^{-1}(u - x)) \cdot \det(\mathbf{L}^{-1}))$  may be computed via automatic differentiation, we can also solve it in closed-form. Firstly, we can observe that

$$\nabla_{\mathbf{L}} \mu(\mathbf{L}^{-1}(u - x)) = \nabla_{\epsilon} \mu(\epsilon) \cdot \nabla_{\mathbf{L}} (\mathbf{L}^{-1} \cdot (u - x)) \quad (50)$$

$$= \nabla_{\mathbf{L}} (\nabla_{\epsilon} \mu(\epsilon) \cdot \mathbf{L}^{-1} \cdot (u - x)) \quad (51)$$

$$= -\mathbf{L}^{-\top} \cdot \nabla_{\epsilon} \mu(\epsilon) \cdot (u - x)^{\top} \cdot \mathbf{L}^{-\top} \quad (52)$$

and

$$\nabla_{\mathbf{L}} \det(\mathbf{L}^{-1}) = -\det(\mathbf{L})^{-1} \cdot \mathbf{L}^{-\top}. \quad (53)$$

We can combine this to resolve it in closed form to:

$$\begin{aligned} \nabla_{\mathbf{L}} (\mu(\mathbf{L}^{-1}(u - x)) \cdot \det(\mathbf{L}^{-1})) &= -\mathbf{L}^{-\top} \cdot \nabla_{\epsilon} \mu(\epsilon) \cdot (u - x)^{\top} \cdot \mathbf{L}^{-\top} \cdot \det(\mathbf{L}^{-1}) \\ &\quad - \mu(\mathbf{L}^{-1}(u - x)) \cdot \det(\mathbf{L})^{-1} \cdot \mathbf{L}^{-\top} \end{aligned} \quad (54)$$

$$\begin{aligned} &= -\mathbf{L}^{-\top} \cdot \nabla_{\epsilon} \mu(\epsilon) \cdot (\mathbf{L}^{-1}(u - x))^{\top} \cdot \det(\mathbf{L}^{-1}) \\ &\quad - \mu(\epsilon) \cdot \det(\mathbf{L})^{-1} \cdot \mathbf{L}^{-\top} \end{aligned} \quad (55)$$

$$\begin{aligned} &= -\mathbf{L}^{-\top} \cdot \nabla_{\epsilon} \mu(\epsilon) \cdot \epsilon^{\top} \cdot \det(\mathbf{L}^{-1}) \\ &\quad - \mu(\epsilon) \cdot \det(\mathbf{L})^{-1} \cdot \mathbf{L}^{-\top} \end{aligned} \quad (56)$$

$$= -\det(\mathbf{L}^{-1}) \cdot (\mathbf{L}^{-\top} \cdot \nabla_{\epsilon} \mu(\epsilon) \cdot \epsilon^{\top} + \mu(\epsilon) \cdot \mathbf{L}^{-\top}). \quad (57)$$

Combing this with equation (49), we have

$$\begin{aligned} & \nabla_{\mathbf{L}} \mathbb{E}_{\epsilon \sim \mu} [f(x + \mathbf{L} \cdot \epsilon)_i] \\ &= \mathbb{E}_{\epsilon \sim \mu} \left[ f(x + \mathbf{L}\epsilon)_i \cdot \nabla_{\mathbf{L}} \left( \mu(\mathbf{L}^{-1}(u - x)) \cdot \det(\mathbf{L}^{-1}) \right) \cdot \det(\mathbf{L}) / \mu(\epsilon) \right] \\ &= \mathbb{E}_{\epsilon \sim \mu} \left[ f(x + \mathbf{L}\epsilon)_i \cdot -\det(\mathbf{L}^{-1}) \cdot (\mathbf{L}^{-\top} \cdot \nabla_{\epsilon} \mu(\epsilon) \cdot \epsilon^{\top} + \mu(\epsilon) \cdot \mathbf{L}^{-\top}) \cdot \det(\mathbf{L}) / \mu(\epsilon) \right] \end{aligned} \quad (58)$$

$$= \mathbb{E}_{\epsilon \sim \mu} \left[ f(x + \mathbf{L}\epsilon)_i \cdot -(\mathbf{L}^{-\top} \cdot \nabla_{\epsilon} \mu(\epsilon) \cdot \epsilon^{\top} + \mu(\epsilon) \cdot \mathbf{L}^{-\top}) / \mu(\epsilon) \right] \quad (59)$$

$$= \mathbb{E}_{\epsilon \sim \mu} \left[ f(x + \mathbf{L}\epsilon)_i \cdot -(\mathbf{L}^{-\top} \cdot \nabla_{\epsilon} \mu(\epsilon) \cdot \epsilon^{\top} / \mu(\epsilon) + \mathbf{L}^{-\top}) \right] \quad (60)$$

$$= \mathbb{E}_{\epsilon \sim \mu} \left[ f(x + \mathbf{L}\epsilon)_i \cdot \mathbf{L}^{-\top} \cdot (-1 + \nabla_{\epsilon} - \log \mu(\epsilon) \cdot \epsilon^{\top}) \right]. \quad (61)$$

□

## B DISCUSSION OF PROPERTIES OF $f$ FOR FINITELY DEFINED $f_{\epsilon}$ AND $\nabla f_{\epsilon}$

When we have a function  $f$  that is not defined with a compact range with  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , and have a density  $\mu$  with unbounded support (e.g., Gaussian or Cauchy), we may experience  $f_{\epsilon}$  or even  $\nabla f_{\epsilon}$  to not be finitely defined. For example, virtually any distribution with full support on  $\mathbb{R}$  leads to the smoothing  $f_{\epsilon}$  of the degenerate function  $f : x \mapsto \exp(\exp(\exp(\exp(x^2))))$  to not be finitely defined.

We say a function, as described via an expectation, is finitely defined iff it is defined (i.e., the expectation has a value) and its value is finite (i.e., not infinity). For example, the first moment of the Cauchy distribution is undefined, and the second moment is infinite; thus, both moments are not finitely defined.

We remark that the considerations in this appendix also apply to prior works that enable the real plane as the output space of  $f$ . We further remark that writing an expression for smoothing and the gradient of an arbitrary function with non-compact range is not necessarily false; however, e.g., any claim that smoothness is guaranteed if the gradient jumps from  $-\infty$  to  $\infty$  (e.g., the power tower in the first paragraph) is not formally correct. We remark that characterizing valid  $f$ s via a Lipschitz or other continuity requirement is not applicable because this would defeat the goal of differentiating non-differentiable and discontinuous  $f$ .

In the following, we discuss when  $f_{\epsilon}$  or  $\nabla f_{\epsilon}$  are finitely defined. For this, let us cover a few preliminaries:

Let a function  $f(x)$  be called  $\mathcal{O}(b(x))$  bounded if there exist  $c, v \in \mathcal{O}(b(x))$  and  $\bar{c}, \bar{v} \in \mathbb{R}$  such that

$$\bar{c} + c(x) \leq f(x) \leq \bar{v} + v(x) \quad \forall x. \quad (62)$$

For example, a function may be called polynomially bounded (wrt. a polynomial  $b(x)$ ) if (but not only if)  $-b(x) \leq f(x) \leq b(x)$ .

Moreover, let a density  $\mu$  with support  $\mathbb{R}$  be called decaying faster than  $b(x)$  if  $\mu(x) \in o(b(x))$ . For example, the standard Gaussian density decays faster than  $\exp(-|x|)$ , i.e.,  $\mu(x) \in o(\exp(-|x|))$ . Additionally, we can say that Gaussian density decays at rate  $\exp(-x^2)$ , i.e.,  $\mu(x) \in \theta(\exp(-x^2))$ .

Now, we can formally characterize finite definedness of  $f_{\epsilon}$  and  $\nabla f_{\epsilon}$ :

**Lemma 8** (Finite Definedness of  $f_{\epsilon}$ ).  *$f_{\epsilon}$  is finitely defined if there exists an increasing function  $b(\cdot)$  such that  $f(x)$  is bounded by  $\mathcal{O}(b(x))$  and  $\mu(\epsilon) \in \mathcal{O}(1/b(\epsilon + \alpha\epsilon)/\epsilon^{(1+\alpha)})$  for some  $\alpha > 0$ .* (63)

*Proof.* To show that  $f_{\epsilon}$  exists, we need to show that

$$\int_{\mathbb{R}} |f(x + \epsilon) \cdot \mu(\epsilon)| d\epsilon \quad (64)$$

is finite for all  $x$ . Let  $\tilde{f}$  be an absolutely upper bound of  $f$ , and w.l.o.g. let us choose  $\tilde{f}(y) = b(y) + \bar{b}$  with  $b(y) > 1$  for  $y \in \mathbb{R}$ . Further, as per the assumptions  $\mu(\epsilon) < \frac{1}{\epsilon^{(1+\alpha)} \cdot b(\epsilon + \alpha\epsilon)} \cdot w$  for all  $\epsilon < \omega_1$  as well as all  $\epsilon > \omega_2$  for some  $w, \omega_1, \omega_2$ . Let us restrict  $\omega_1, \omega_2$  to  $\omega_1 < -|x|/\alpha$  and  $\omega_2 > |x|/\alpha$ . It is trivial to see that

$$\int_{\omega_1}^{\omega_2} |f(x + \epsilon) \cdot \mu(\epsilon)| d\epsilon < \infty. \quad (65)$$

W.l.o.g., let us consider the upper remainder:

$$\int_{\omega_2}^{\infty} |f(x + \epsilon) \cdot \mu(\epsilon)| d\epsilon \leq \int_{\omega_2}^{\infty} |\tilde{f}(x + \epsilon) \cdot \mu(\epsilon)| d\epsilon \quad (66)$$

$$\leq \int_{\omega_2}^{\infty} \left| (b(x + \epsilon) + \bar{b}) \cdot \frac{1}{\epsilon^{(1+\alpha)} \cdot b(\epsilon + \alpha\epsilon)} \cdot w \right| d\epsilon \quad (67)$$

$$= \int_{\omega_2}^{\infty} \left| \left( \frac{b(x + \epsilon)}{\epsilon^{(1+\alpha)} \cdot b(\epsilon + \alpha\epsilon)} + \frac{\bar{b}}{\epsilon^{(1+\alpha)} \cdot b(\epsilon + \alpha\epsilon)} \right) \cdot w \right| d\epsilon \quad (68)$$

$$\leq \int_{\omega_2}^{\infty} \left| \left( \frac{b(x + \epsilon)}{\epsilon^{(1+\alpha)} \cdot b(\epsilon + |x|)} + \frac{\bar{b}}{\epsilon^{(1+\alpha)} \cdot b(\epsilon + \alpha\epsilon)} \right) \cdot w \right| d\epsilon \quad (69)$$

$$\leq \int_{\omega_2}^{\infty} \left| \left( \frac{1}{\epsilon^{(1+\alpha)}} + \frac{\bar{b}}{\epsilon^{(1+\alpha)} \cdot b(\epsilon + \alpha\epsilon)} \right) \cdot w \right| d\epsilon \quad (70)$$

$$< \int_{\omega_2}^{\infty} \left| \frac{1}{\epsilon^{(1+\alpha)}} + \frac{\bar{b}}{\epsilon^{(1+\alpha)}} \right| d\epsilon \cdot w \quad (71)$$

$$= \int_{\omega_2}^{\infty} \left| \frac{1}{\epsilon^{(1+\alpha)}} \right| d\epsilon \cdot w \cdot (1 + \bar{b}) < \infty. \quad (72)$$

That  $\int_{\omega_2}^{\infty} \frac{1}{\epsilon^{(1+\alpha)}} d\epsilon$  is finite for the step in (72) can be shown via

$$\int_{\omega_2}^{\infty} \frac{1}{\epsilon^{(1+\alpha)}} d\epsilon = \int_{\omega_2}^{\infty} \epsilon^{-1-\alpha} d\epsilon = \left[ -\frac{1}{\alpha} \epsilon^{-\alpha} \right]_{\omega_2}^{\infty} = \left[ -\frac{1}{\alpha} \lim_{\epsilon \rightarrow \infty} \epsilon^{-\alpha} + \frac{1}{\alpha} \omega_2^{-\alpha} \right] = \frac{1}{\alpha} \omega_2^{-\alpha}.$$

The same can be shown analogously for the integral  $\int_{-\infty}^{\omega_1}$ . This completes the proof.  $\square$

**Lemma 9** (Finite Definedness of  $\nabla f_\epsilon$ ).  $\nabla f_\epsilon$  is finitely defined if there exists an increasing function  $b(\cdot)$  such that

$$f(x) \text{ is bounded by } \mathcal{O}(b(x)) \quad \text{and} \quad |\mu(\epsilon) \cdot \nabla_\epsilon - \log \mu(\epsilon)| \in \mathcal{O}(1/b(\epsilon + \alpha\epsilon)/\epsilon^{(1+\alpha)}) \quad (73)$$

for some  $\alpha > 0$ .

*Proof.* The proof of Lemma 8 also applies here, but with  $|\mu(\epsilon) \cdot \nabla_\epsilon - \log \mu(\epsilon)| < \frac{1}{\epsilon^{(1+\alpha)} \cdot b(\epsilon + \alpha\epsilon)} \cdot w$  for all  $\epsilon < \omega_1$  as well as all  $\epsilon > \omega_2$  for some  $w, \omega_1, \omega_2$ .  $\square$

**Example 10** (Cauchy and the Identity). Let  $\mu$  be the density of a Cauchy distribution and let  $f(x) = x$ . The tightest  $b$  for  $f(x) \in \mathcal{O}(b(x))$  is  $b(x) = x$ .

We have  $\mu(\epsilon) \in \theta(1/\epsilon^2)$  and thus  $\mu(\epsilon) \notin o(1/\epsilon^2)$ .  $f_\epsilon$ , i.e., the mean of the Cauchy distribution is not defined.

However, its gradient  $\nabla f_\epsilon = 1$  is indeed finitely defined. In particular, we can see that

$$\mu(\epsilon) \cdot \nabla_\epsilon - \log \mu(\epsilon) = \frac{2\epsilon}{\pi \cdot (1 + \epsilon^2) \cdot (1 + \epsilon^2)} \in \theta(1/\epsilon^3). \quad (74)$$

This is an intriguing property of the Cauchy distribution (or other edge cases) where  $f_\epsilon$  is undefined whereas  $\nabla f_\epsilon$  is finitely and well-defined. In practice, we often only require the gradient for stochastic gradient descent, which means that we often only require  $\nabla f_\epsilon$  to be well defined and do not necessarily need to evaluate  $f_\epsilon$  depending on the application.

Additional discussions for the Cauchy distribution and an extension of stochastic smoothing to the  $k$ -sample median can be found in the next appendix.

## C STOCHASTIC SMOOTHING, MEDIAN, AND THE CAUCHY DISTRIBUTION

In this section, we provide a discussion of a special case of stochastic smoothing with the Cauchy distribution, and provide an extension of stochastic smoothing to the  $k$ -sample median. This becomes important if the range of  $f$  is not subset of a compact set, and thus  $\mathbb{E}_{\epsilon \sim \mu} [f(x + \epsilon)]$  becomes undefined for some choice of distribution  $\mu$ . For example, for  $f(x + \epsilon) = \epsilon$  and  $\mu$  being the density of a Cauchy distribution,  $\mathbb{E}_{\epsilon \sim \mu} [f(x + \epsilon)] = \mathbb{E}_{\epsilon \sim \mu} [\epsilon]$  is undefined. Nevertheless, even in this case, the gradient estimators discussed in this paper for  $\nabla_x \mathbb{E}_{\epsilon \sim \mu} [f(x + \epsilon)]$  remain well defined. This is practically relevant because  $\mathbb{E}_{\epsilon \sim \mu} [f(x + \epsilon)]$  does not need to be finitely defined as long as  $\nabla_x \mathbb{E}_{\epsilon \sim \mu} [f(x + \epsilon)]$  is well defined. Further, we remark that the undefinedness of  $\mathbb{E}_{\epsilon \sim \mu} [f(x + \epsilon)]$  requires the range of  $f$  to be unbounded, i.e., if there exists a maximum / minimum possible output, then it is well defined. Moreover, there exist  $f$  with unbounded range for which  $\mathbb{E}_{\epsilon \sim \mu} [f(x + \epsilon)]$  also remains well defined.

To account for cases where  $\mathbb{E}_{\epsilon \sim \mu} [f(x + \epsilon)]$  may not be well defined or not a robust statistic, we introduce an extension of smoothing to the median. We begin by defining the  $k$ -sample median.

**Definition 11** ( $k$ -Sample Median). For a number of samples  $k > 1$ , and a distribution  $\zeta$ , we say that

$$\mathbb{E}_{z_1, z_2, \dots, z_k \sim \zeta} \left[ \text{median} \{z_1, z_2, \dots, z_k\} \right] \quad (75)$$

is the  $k$ -sample median. For multivariate distributions, let median be the per-dimension median.

Indeed, for  $k \geq 5$ , the  $k$ -sample median estimator is shown to have finite variance for the Cauchy distribution (Theorem 3 and Example 2 in [56]), which implies a well defined  $k$ -sample median. Moreover, for any distribution with a density of the median bounded away from 0, the first and second moments are guaranteed to be finitely defined for sufficiently large  $k$ . This is important for non-trivial  $f$  with  $f(\epsilon) \neq \epsilon$  for at least one  $\epsilon$  with  $\epsilon \sim \mu$ , which implies  $\zeta \neq \mu$ . Thus, rather than computing and differentiating the expected value, we can differentiate the  $k$ -sample median.

**Lemma 12** (Differentiation of the  $k$ -Sample Median). *With the  $k$ -sample median smoothing as*

$$f_\epsilon^{(k)}(x) = \mathbb{E}_{\epsilon_1, \dots, \epsilon_k \sim \mu} \left[ \text{median} \{f(x + \epsilon_1), \dots, f(x + \epsilon_k)\} \right], \quad (76)$$

*we can differentiate  $f_\epsilon^{(k)}(x)$  as*

$$\nabla_x f_\epsilon^{(k)}(x) = \mathbb{E}_{\epsilon_1, \dots, \epsilon_k \sim \mu} \left[ f(x + \epsilon_{r(\epsilon)}) \cdot \nabla_{\epsilon_{r(\epsilon)}} - \log \mu(\epsilon_{r(\epsilon)}) \right] \quad (77)$$

*where  $r(\epsilon)$  is the arg-median of the set  $\{f(x + \epsilon_1), \dots, f(x + \epsilon_k)\}$ , which is equivalent to the implicit definition via  $f(x + \epsilon_{r(\epsilon)}) = \text{median} \{f(x + \epsilon_1), \dots, f(x + \epsilon_k)\}$ .*

*Proof.* We denote  $\epsilon_{1:k} \sim \mu^{(1:k)}$  such that  $\epsilon_{1:k} = [\epsilon_1^\top, \dots, \epsilon_k^\top]^\top$  and  $\epsilon_i \sim \mu \forall i \in \{1, \dots, k\}$ .

$$\nabla_x f_\epsilon^{(k)}(x) = \nabla_x \mathbb{E}_{\epsilon_1, \dots, \epsilon_k \sim \mu} \left[ \text{median} \{f(x + \epsilon_1), \dots, f(x + \epsilon_k)\} \right] \quad (78)$$

$$= \nabla_x \mathbb{E}_{\epsilon_{1:k} \sim \mu^{(1:k)}} \left[ \text{median} \{f(x + \epsilon_1), \dots, f(x + \epsilon_k)\} \right] \quad (79)$$

$$= \nabla_x \int_{\mathbb{R}^{n \cdot k}} \text{median} \{f(x + \epsilon_1), \dots, f(x + \epsilon_k)\} \cdot \mu^{(1:k)}(\epsilon_{1:k}) d\epsilon_{1:k} \quad (80)$$

$$(x_1, \dots, x_k = x) = \sum_{j=1}^k \nabla_{x_j} \int_{\mathbb{R}^{n \cdot k}} \text{median} \{f(x_1 + \epsilon_1), \dots, f(x_k + \epsilon_k)\} \cdot \mu^{(1:k)}(\epsilon_{1:k}) d\epsilon_{1:k} \quad (81)$$

As a shorthand, we abbreviate the indicator  $\mathbb{1}_{f(x_j+\epsilon_j)=\text{median}\{f(x_1+\epsilon_1),\dots,f(x_k+\epsilon_k)\}}$  as  $\mathbb{1}_{j,\epsilon_{1:k}}$  and abbreviate  $\mathbb{1}_{f(u_j)=\text{median}\{f(u_1),\dots,f(u_k)\}}$  as  $\mathbb{1}_{j,u_{1:k}}$ :

$$\nabla_x f_\epsilon^{(k)}(x) = \sum_{j=1}^k \nabla_{x_j} \int_{\mathbb{R}^{n \cdot k}} f(x_j + \epsilon_j) \cdot \mathbb{1}_{j,\epsilon_{1:k}} \cdot \mu^{(1:k)}(\epsilon_{1:k}) d\epsilon_{1:k} \quad (82)$$

$$= \sum_{j=1}^k \nabla_{x_j} \int_{\mathbb{R}^{n \cdot k}} f(u) \cdot \mathbb{1}_{j,u_{1:k}} \cdot \mu^{(1:k)}(u_{1:k} - x) du_{1:k} \quad (83)$$

$$= \sum_{j=1}^k \int_{\mathbb{R}^{n \cdot k}} f(u) \cdot \mathbb{1}_{j,u_{1:k}} \cdot \nabla_{x_j} \mu^{(1:k)}(u_{1:k} - x) du_{1:k} \quad (84)$$

$$= \sum_{j=1}^k \int_{\mathbb{R}^{n \cdot k}} f(x + \epsilon_j) \cdot \mathbb{1}_{j,\epsilon_{1:k}} \cdot -\nabla_{\epsilon_j} \mu^{(1:k)}(\epsilon_{1:k}) d\epsilon_{1:k} \quad (85)$$

We have

$$\nabla_{\epsilon_j} \mu^{(1:k)}(\epsilon_{1:k}) = \mu^{(1:k)}(\epsilon_{1:k}) \cdot \nabla_{\epsilon_j} \log \mu^{(1:k)}(\epsilon_{1:k}) = \mu^{(1:k)}(\epsilon_{1:k}) \cdot \nabla_{\epsilon_j} \log \mu(\epsilon_j). \quad (86)$$

Thus,

$$\nabla_x f_\epsilon^{(k)}(x) = \sum_{j=1}^k \int_{\mathbb{R}^{n \cdot k}} f(x + \epsilon_j) \cdot \mathbb{1}_{j,\epsilon_{1:k}} \cdot -\mu^{(1:k)}(\epsilon_{1:k}) \cdot \nabla_{\epsilon_j} \log \mu(\epsilon_j) d\epsilon_{1:k} \quad (87)$$

$$= \int_{\mathbb{R}^{n \cdot k}} \sum_{j=1}^k [\mathbb{1}_{j,\epsilon_{1:k}} \cdot f(x + \epsilon_j) \cdot \nabla_{\epsilon_j} - \log \mu(\epsilon_j)] \cdot \mu^{(1:k)}(\epsilon_{1:k}) d\epsilon_{1:k} \quad (88)$$

Indicating the choice of median in dependence of  $\epsilon_{1:k}$ , we define  $r(\epsilon_{1:k})$  s.t.  $\mathbb{1}_{r(\epsilon_{1:k}),\epsilon_{1:k}} = 1$ . Thus,

$$\nabla_x f_\epsilon^{(k)}(x) = \int_{\mathbb{R}^{n \cdot k}} f(x + \epsilon_{r(\epsilon_{1:k})}) \cdot \nabla_{\epsilon_{r(\epsilon_{1:k})}} - \log \mu(\epsilon_{r(\epsilon_{1:k})}) \cdot \mu^{(1:k)}(\epsilon_{1:k}) d\epsilon_{1:k} \quad (89)$$

$$= \mathbb{E}_{\epsilon_{1:k} \sim \mu^{(1:k)}} \left[ f(x + \epsilon_{r(\epsilon_{1:k})}) \cdot \nabla_{\epsilon_{r(\epsilon_{1:k})}} - \log \mu(\epsilon_{r(\epsilon_{1:k})}) \right] \quad (90)$$

This concludes the proof.  $\square$

Empirically, we can estimate  $\nabla_x f_\epsilon^{(k)}(x)$  for  $s$  propagated samples ( $s > k$ ) without bias as

$$\nabla_x f_\epsilon^{(k)}(x) \triangleq \sum_{i=1}^s \left[ q_i \cdot f(x + \epsilon_i) \cdot \nabla_{\epsilon_i} - \log \mu(\epsilon_i) \right] \quad \epsilon_1, \dots, \epsilon_s \sim \mu \quad (91)$$

where  $q_i$  is the probability of  $f(x + \epsilon_i)$  being the median in a subset of  $k$  samples, i.e., under uniqueness of  $g_i$ s, we have

$$q_i = \frac{\sum_{\{h_1, \dots, h_k\} \subset \{g_1, \dots, g_s\}} \mathbb{1}(g_i = \text{median}\{h_1, \dots, h_k\})}{\binom{s}{k}} \quad g_i := f(x + \epsilon_i). \quad (92)$$

We remark that, in case of non-uniqueness, it is adequate to split the probability among the candidates; however, under non-discreteness assumptions on  $f$  (density of  $\zeta < \infty$ , the converse typically implies the range of  $f$  being a subset of a compact set), this almost surely (with probability 1) does not occur.

We have shown that the  $k$ -sample median  $f_\epsilon^{(k)}(x)$  is differentiable and demonstrated an unbiased gradient estimator for it. A straightforward extension for the case of  $f$  being differentiable is differentiating through the median via a  $k \rightarrow \infty$ -sample median, e.g., via setting  $s = k^2$ . The  $k \rightarrow \infty$  extension for differentiating through the median itself requires  $f$  being differentiable because, for discontinuous  $f$ ,  $f_\epsilon^{(k)}(x)$  is differentiable only for  $k < \infty$ . (As an illustration, the median of the Heaviside function under a symmetric perturbation  $\mu$  with density at 0 bounded away from 0 is the exactly the Heaviside function.)

## D EXPERIMENTAL DETAILS

**MNIST Sorting Benchmark Experiments** We train for 100 000 steps at a learning rate of 0.001 with the Adam optimizer using a batch size of 100. Following the requirements of the benchmark, we use the same model as previous works [7], [8], [11]. That is, two convolutional layers with a kernel size of  $5 \times 5$ , 32 and 64 channels respectively, each followed by a ReLU and MaxPool layer; after flattening, this is followed by a fully connected layer with a size of 64, a ReLU layer, and a fully connected output layer mapping to a scalar. For each distribution and number of samples, we choose the optimal  $\gamma \in \{1, 1/3, 0.1\}$ .

**Warcraft Shortest-Path Benchmark Experiments** Following the established protocol [17], we train for 50 epochs with the Adam optimizer at a batch size of 70 and an initial learning rate of 0.001. The learning rate decays by a factor of 10 after 30 and 40 epochs each. The model is the first block of ResNet18. The hyperparameter  $\gamma = 1/\beta$  as specified in Figures 13 and 14.

**Utah Teapot Camera Pose Optimization Experiments** We initialize the pose to be perturbed by angles uniformly sampled from  $[15^\circ, 75^\circ]$ . The ground truth orientation is randomly sampled from the sphere of possible orientations. The ground truth camera angle is  $20^\circ$ , and the ground truth camera distance is uniformly sampled from  $[2.5, 4]$ . The initial camera distance is sampled as being uniformly offset by  $[-0.5, 6]$ , thus the feasible set of initial camera distance guesses lies in  $[2, 10]$ . The initial camera angle is uniformly sampled from  $[10^\circ, 30^\circ]$ . We optimize for 1 000 steps with the Adam optimizer  $[(\beta_1, \beta_2) = (0.5, 0.99)]$  and the CosineAnnealingLR scheduler with an initial learning rate of 0.3. We schedule the diagonal of  $\mathbf{L}$  to decay exponentially from  $[0.1, 5^\circ, 5^\circ, 0.25^\circ] \cdot 10^{0.75}$  to  $[0.1, 5^\circ, 5^\circ, 0.25^\circ] \cdot 10^{-1.75}$  (the dimensions are camera distance, 2 pose angles, and the camera angle). As discussed, the success criterion is finding the angle within  $5^\circ$  of the ground truth angle. There is typically no local minimum within  $5^\circ$  and it is a reliable indicator for successful alignment.

**Differentiable Cryo-Electron Tomography Experiments** The ground truth values of the parameters are set to 300 kV for acceleration voltage, 3 mm for the focal length, and the ground truth sample specimen is centered as  $(x, y) = (0, 0)$  nm units. For reporting errors, the acceleration voltages are normalized by a factor of 100 to ensure that all parameters vary over commensurate ranges. For the 2-parameter optimization, the feasible set of acceleration voltage varied over a range of  $[0, 1000]$  kV and the feasible set of the specimen’s  $x$ -position varied over the range  $[-5, 5]$ . For the 4-parameter optimization, the feasible set of acceleration voltage varied over a range of  $[0, 600]$  kV, the focal length ranges over  $[0, 6]$  mm, the  $x$ - and  $y$ -positions range over  $[-3, 3]$ . We use the Adam optimizer for both experiments, with  $[(\beta_1, \beta_2) = (0.5, 0.9)]$ . For the MC Search baseline, we generate sets of  $n$  uniform random points in the feasible region of the parameters, generate micrographs for these random parameter tuples using the TEM simulator [53], and identify the parameter tuple in the set having the lowest mean squared error with respect to the ground truth image. The RMSE between this parameter tuple and the ground truth parameters is the metric for the specific set of  $n$  randomly generated values. This is repeated 20 times to obtain the mean and standard deviation of the RMSE metric at that  $n$ .

### D.1 ASSETS

*List of assets:*

- The sixth platonic solid (aka. Teapotahedron or Utah tea pot) [57] [License N/A]
- Multi-digit MNIST [8], which builds on MNIST [58] [MIT License / CC License]
- Warcraft shortest-path data set [17] [MIT License]
- PyTorch [59] [BSD 3-Clause License]
- TEM-simulator [53] [GNU General Public License]

### D.2 RUNTIMES

The runtimes for sorting and shortest-path experiments are for one full training on 1 GPU. The pose optimization experiment runtimes are the total time for all 768 seeds on 1 GPU. For the TEM-



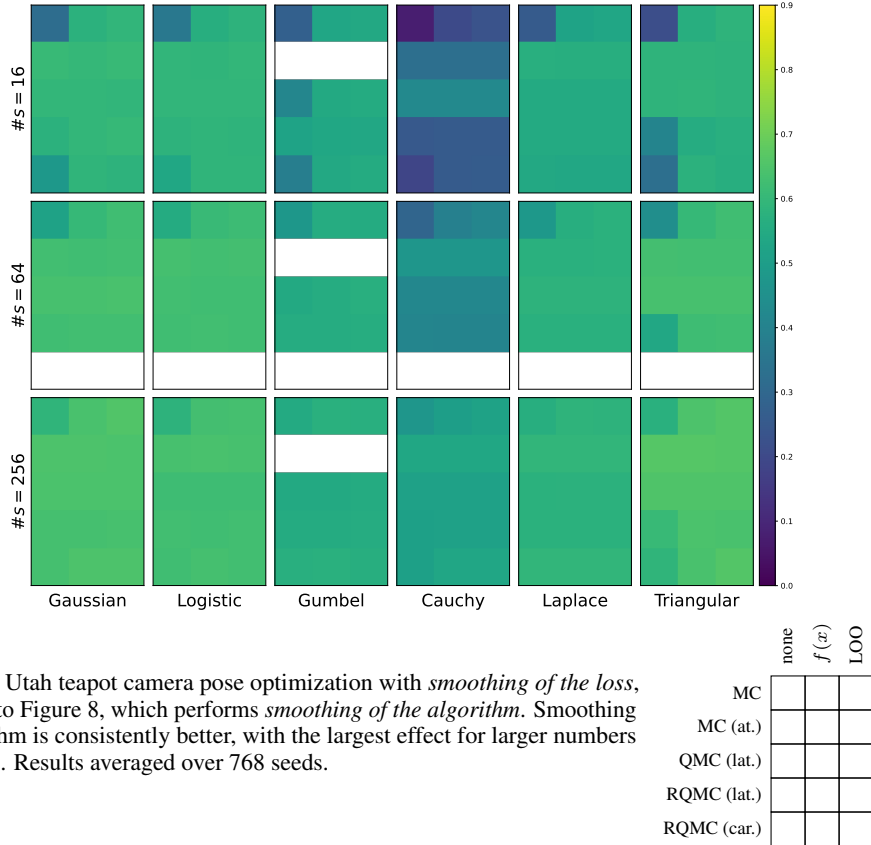
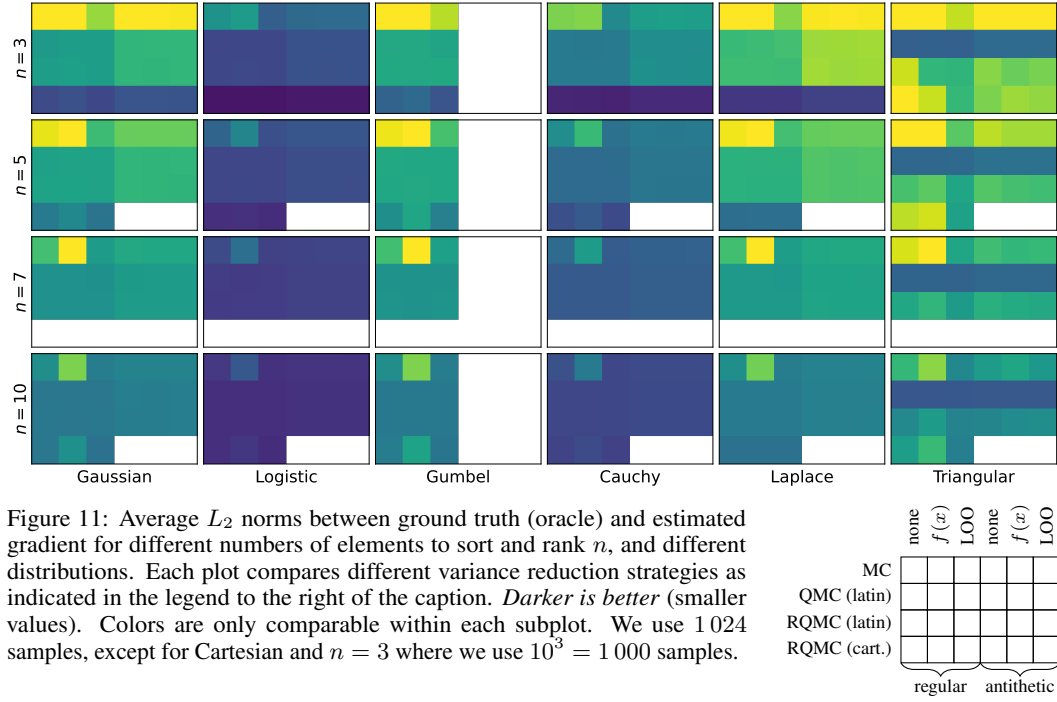
simulator, we report the CPU time per simulation sample, which is the dominant and only the measurable component of the total optimization routine time. The choice of distribution, covariate, and choice of variance reduction does not have a measurable effect on training times.

- MNIST Sorting Benchmark Experiments [1 Nvidia V100 GPU]
  - Training w/ 256 samples: 65 min
  - Training w/ 1 024 samples: 67 min
  - Training w/ 2 048 samples: 68 min
  - Training w/ 8 192 samples: 77 min
  - Training w/ 32 768 samples: 118 min
- Warcraft Shortest-Path Benchmark Experiments [1 Nvidia V100 GPU]
  - Training w/ 10 samples: 9 min
  - Training w/ 100 samples: 19 min
  - Training w/ 1 000 samples: 26 min
  - Training w/ 10 000 samples: 101 min
- Utah Teapot Camera Pose Optimization Experiments [1 Nvidia A6000 GPU]
  - Optimization on 768 seeds w/ 16 samples: 25 min
  - Optimization on 768 seeds w/ 64 samples: 81 min
  - Optimization on 768 seeds w/ 256 samples: 362 min
- Differentiable Cryo-Electron Tomography Experiments [CPU: 44 Intel Xeon Gold 5118]
  - Simulator time per sample on 1 CPU core: 67 sec

## E ADDITIONAL EXPERIMENTAL RESULTS

Table 3: Extension of Table 2 with additional numbers of samples and standard deviations.

Baselines		Neu.S.	Soft.S.	L. DSN	C. DSN	E. DSN	OT. S.
—		71.3	70.7	77.2	84.9	85.0	81.1
Sampling	#s	Gauss.	Logis.	Gumbel	Cauchy	Laplace	Trian.
vanilla	256	82.3±2.0	82.8±0.9	79.2±9.7	68.1±19.3	82.6±0.8	81.3±1.2
best (cv)	256	83.1±1.6	82.7±1.8	81.6±3.6	55.6±13.3	83.7±0.8	82.7±1.1
vanilla	1024	81.3±9.1	83.7±0.7	82.0±1.6	68.5±24.8	80.6±9.0	82.8±1.0
best (cv)	1024	83.9±0.6	84.0±0.5	84.2±0.6	73.0±12.6	84.3±0.6	82.4±1.6
vanilla	2048	84.1±0.6	83.6±0.8	84.0±0.5	75.7±11.6	83.8±0.7	83.2±0.6
best (cv)	2048	84.2±0.5	84.2±0.6	84.6±0.4	82.0±2.2	84.8±0.5	83.4±0.5
vanilla	8192	84.0±0.6	84.2±0.8	84.0±0.6	83.6±1.0	83.9±1.0	83.6±0.7
best (cv)	8192	84.4±0.6	84.5±0.5	84.1±0.7	84.3±0.5	84.3±0.4	83.7±0.4
vanilla	32768	84.2±0.5	84.1±0.4	84.5±0.7	84.9±0.5	84.4±0.5	83.4±0.8
best (cv)	32768	84.4±0.4	84.4±0.4	84.8±0.5	85.1±0.4	84.4±0.4	84.0±0.3



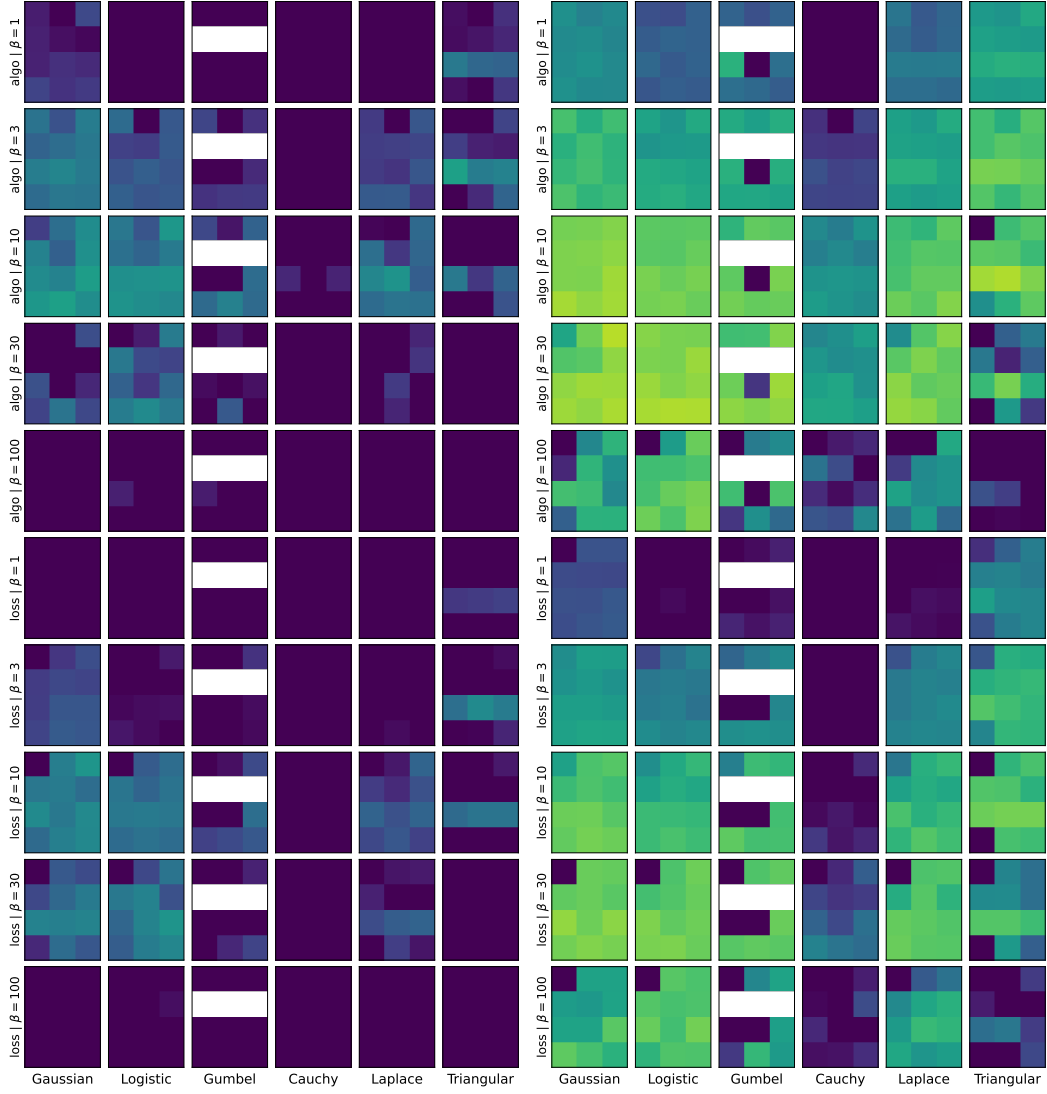
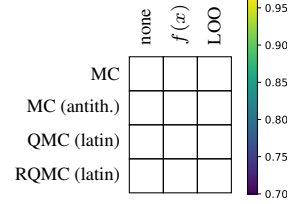


Figure 13: Warcraft shortest-path experiment. *Left: 10 samples. Right: 100 samples.* Averaged over 5 seeds. *Brighter is better.* Values between subplots are comparable. The displayed range is [70%, 96.5%].



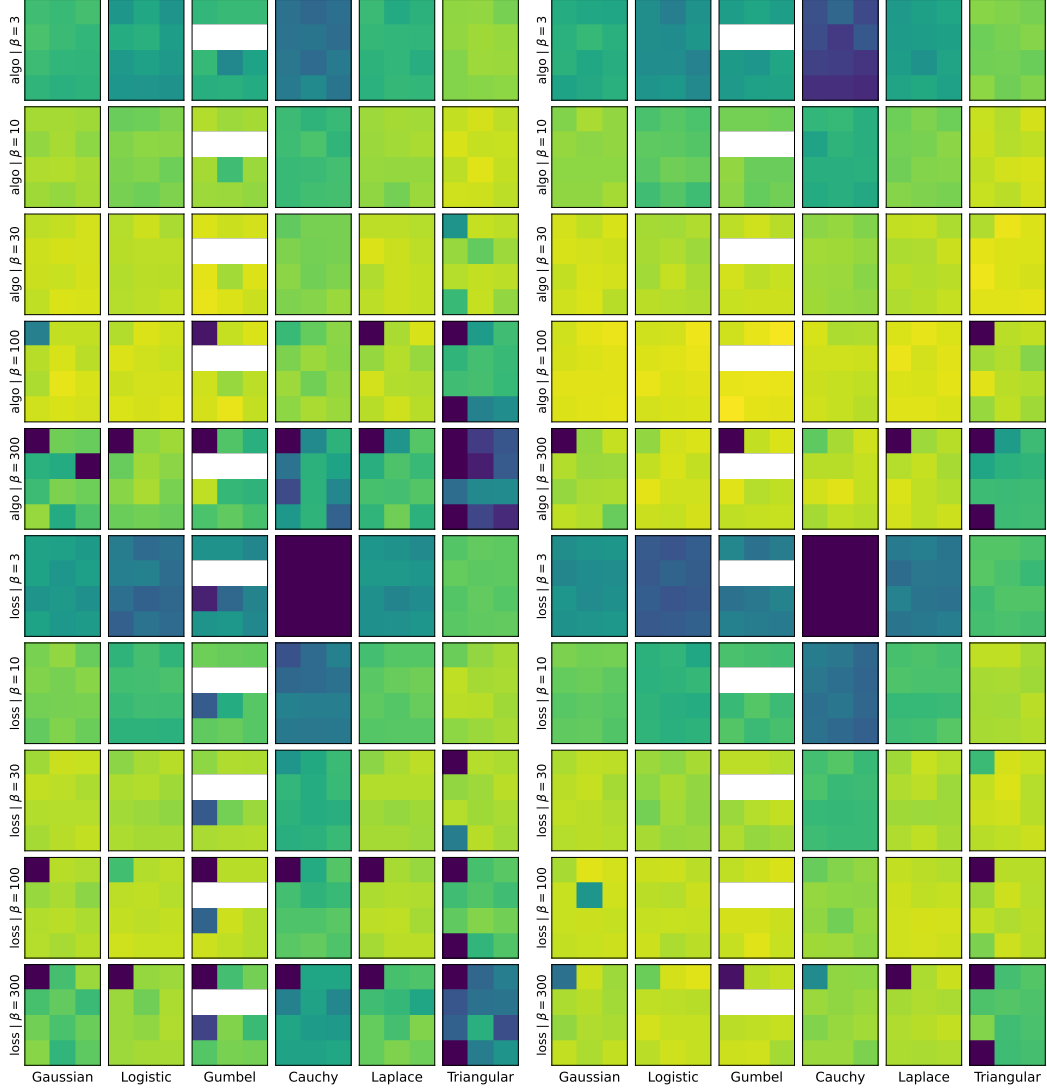
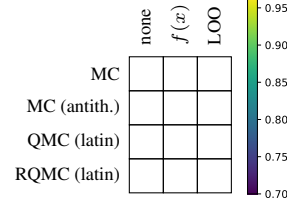


Figure 14: Warcraft shortest-path experiment. *Left: 1000 samples. Right: 10000 samples.* Averaged over 5 seeds. *Brighter is better.* Values between subplots are comparable. The displayed range is [70%, 96.5%].



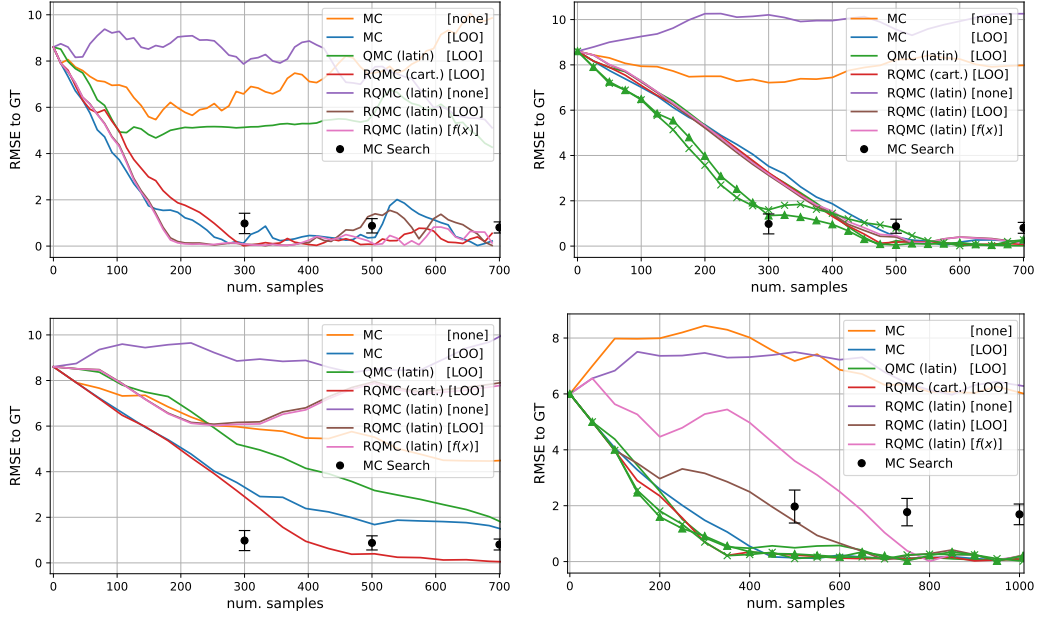


Figure 15: Cryo-Electron Tomography Experiments: RMSE with respect to Ground Truth parameters for different number of parameters optimized and for different number of samples per optimization step: (Top Left) 2-parameters & number of samples=9, (Top Right) 2-parameters & number of samples=25, (Bottom Left) 2-parameters & number of samples=36, (Bottom Right) 4-parameters. No marker lines correspond to Gaussian,  $\times$  corresponds to Laplace, and  $\triangle$  corresponds to Triangular distributions. Ascertaining optimal parameters with minimal evaluations is important not just for high resolution imaging, but also to minimize radiation damage to the specimen. In this light, of the covariate choices, LOO generally leads to best improvement and *none* consistently leads to deterioration in performance. The Laplace and Triangular distributions lead to best performance. For the Gaussian distribution, Cartesian RQMC is generally exhibiting best results.

Table 4: Individual absolute values from the variance simulations for differentiable sorting in Figure 3. The minimum and values within 1% of the minimum are indicated as bold.

(a) values for Gaussian ( $n = 3$ )							(b) values for Gaussian ( $n = 5$ )						
	none	$f(x)$	LOO	none	$f(x)$	LOO		none	$f(x)$	LOO	none	$f(x)$	LOO
	regular			antithetic				regular			antithetic		
MC	0.0084	0.0079	0.0046	0.0055	0.0054	0.0053	MC	0.0241	0.0308	0.0171	0.0192	0.0192	0.0192
QMC (lat.)	0.0029	0.0030	0.0030	0.0036	0.0036	0.0036	QMC (lat.)	0.0143	0.0144	0.0144	0.0164	0.0164	0.0164
RQMC (l.)	0.0030	0.0030	0.0030	0.0036	0.0035	0.0036	RQMC (l.)	0.0145	0.0145	0.0144	0.0164	0.0164	0.0162
RQMC (c.)	0.0012	0.0013	<b>0.0012</b>	0.0014	0.0014	0.0014	RQMC (c.)	0.0103	0.0116	<b>0.0097</b>	—	—	—
(c) values for Logistic ( $n = 3$ )							(d) values for Logistic ( $n = 5$ )						
	none	$f(x)$	LOO	none	$f(x)$	LOO		none	$f(x)$	LOO	none	$f(x)$	LOO
	regular			antithetic				regular			antithetic		
MC	0.0028	0.0030	0.0016	0.0019	0.0019	0.0019	MC	0.0081	0.0114	0.0061	0.0067	0.0067	0.0067
QMC (lat.)	0.0012	0.0012	0.0012	0.0014	0.0014	0.0014	QMC (lat.)	0.0053	0.0053	0.0054	0.0060	0.0060	0.0060
RQMC (l.)	0.0012	0.0012	0.0012	0.0014	0.0013	0.0014	RQMC (l.)	0.0053	0.0054	0.0053	0.0060	0.0060	0.0059
RQMC (c.)	<b>0.0003</b>	0.0003	<b>0.0003</b>	0.0004	0.0004	0.0004	RQMC (c.)	0.0033	0.0036	<b>0.0033</b>	—	—	—
(e) values for Gumbel ( $n = 3$ )							(f) values for Gumbel ( $n = 5$ )						
	none	$f(x)$	LOO	none	$f(x)$	LOO		none	$f(x)$	LOO	none	$f(x)$	LOO
	regular			antithetic				regular			antithetic		
MC	0.0086	0.0082	0.0048	—	—	—	MC	0.0243	0.0323	0.0177	—	—	—
QMC (lat.)	0.0033	0.0033	0.0032	—	—	—	QMC (lat.)	0.0151	0.0149	0.0150	—	—	—
RQMC (l.)	0.0033	0.0033	0.0033	—	—	—	RQMC (l.)	0.0150	0.0151	0.0150	—	—	—
RQMC (c.)	0.0017	0.0018	<b>0.0014</b>	—	—	—	RQMC (c.)	0.0124	0.0148	<b>0.0109</b>	—	—	—
(g) values for Cauchy ( $n = 3$ )							(h) values for Cauchy ( $n = 5$ )						
	none	$f(x)$	LOO	none	$f(x)$	LOO		none	$f(x)$	LOO	none	$f(x)$	LOO
	regular			antithetic				regular			antithetic		
MC	0.0043	0.0044	0.0026	0.0030	0.0030	0.0030	MC	0.0123	0.0169	0.0094	0.0102	0.0101	0.0102
QMC (lat.)	0.0022	0.0022	0.0022	0.0027	0.0027	0.0027	QMC (lat.)	0.0088	0.0087	0.0088	0.0098	0.0098	0.0098
RQMC (l.)	0.0022	0.0022	0.0022	0.0027	0.0026	0.0027	RQMC (l.)	0.0088	0.0088	0.0087	0.0098	0.0097	0.0097
RQMC (c.)	0.0006	0.0006	<b>0.0005</b>	0.0006	0.0006	0.0006	RQMC (c.)	0.0061	0.0070	<b>0.0056</b>	—	—	—
(i) values for Laplace ( $n = 3$ )							(j) values for Laplace ( $n = 5$ )						
	none	$f(x)$	LOO	none	$f(x)$	LOO		none	$f(x)$	LOO	none	$f(x)$	LOO
	regular			antithetic				regular			antithetic		
MC	0.0086	0.0074	0.0044	0.0054	0.0054	0.0054	MC	0.0245	0.0305	0.0176	0.0191	0.0192	0.0192
QMC (lat.)	0.0037	0.0037	0.0038	0.0046	0.0046	0.0047	QMC (lat.)	0.0159	0.0160	0.0160	0.0182	0.0180	0.0182
RQMC (l.)	0.0037	0.0037	0.0037	0.0047	0.0046	0.0046	RQMC (l.)	0.0160	0.0159	0.0159	0.0182	0.0181	0.0181
RQMC (c.)	<b>0.0009</b>	<b>0.0009</b>	<b>0.0009</b>	0.0010	0.0011	0.0010	RQMC (c.)	<b>0.0091</b>	<b>0.0091</b>	<b>0.0091</b>	—	—	—
(k) values for Triangular ( $n = 3$ )							(l) values for Triangular ( $n = 5$ )						
	none	$f(x)$	LOO	none	$f(x)$	LOO		none	$f(x)$	LOO	none	$f(x)$	LOO
	regular			antithetic				regular			antithetic		
MC	0.1191	0.0683	0.0490	0.0659	0.0624	0.0602	MC	0.3329	0.2779	0.1857	0.2255	0.2157	0.2149
QMC (lat.)	<b>0.0166</b>	0.0169	<b>0.0166</b>	0.0189	0.0188	0.0188	QMC (lat.)	<b>0.0844</b>	<b>0.0845</b>	<b>0.0851</b>	0.0932	0.0931	0.0928
RQMC (l.)	0.0498	0.0358	0.0352	0.0444	0.0417	0.0431	RQMC (l.)	0.1768	0.1872	0.1479	0.1827	0.1765	0.1737
RQMC (c.)	0.0682	0.0494	0.0361	0.0435	0.0461	0.0452	RQMC (c.)	0.2251	0.2325	0.1430	—	—	—

Table 5: Individual absolute values from the variance simulations for differentiable shortest-paths in Figure 4. The minimum and values within 1% of the minimum are indicated as bold.

(a) values for Gaussian ( $8 \times 8$ )						
	none $f(x)$ LOO			none $f(x)$ LOO		
	regular			antithetic		
MC	1330.01	4.17	4.17	8.32	8.32	8.34
QMC (lat.)	<b>4.04</b>	<b>4.04</b>	<b>4.04</b>	8.04	8.04	8.07
RQMC (l.)	4.25	<b>4.05</b>	<b>4.05</b>	8.10	8.09	8.12
(c) values for Logistic ( $8 \times 8$ )						
	none $f(x)$ LOO			none $f(x)$ LOO		
	regular			antithetic		
MC	1449.44	4.53	4.53	9.04	9.04	9.05
QMC (lat.)	<b>4.42</b>	<b>4.42</b>	<b>4.43</b>	8.80	8.80	8.83
RQMC (l.)	<b>4.44</b>	<b>4.44</b>	<b>4.44</b>	8.88	8.87	8.90
(e) values for Gumbel ( $8 \times 8$ )						
	none $f(x)$ LOO			none $f(x)$ LOO		
	regular			antithetic		
MC	2275.31	10.35	9.08	—	—	—
QMC (lat.)	9.11	<b>8.84</b>	<b>8.85</b>	—	—	—
RQMC (l.)	11.33	<b>8.91</b>	<b>8.91</b>	—	—	—
(b) values for Gaussian ( $12 \times 12$ )						
	none $f(x)$ LOO			none $f(x)$ LOO		
	regular			antithetic		
MC	6800.98	20.93	20.95	41.82	41.78	41.88
QMC (lat.)	<b>20.60</b>	<b>20.60</b>	<b>20.65</b>	41.12	41.11	41.18
RQMC (l.)	21.69	<b>20.66</b>	<b>20.68</b>	41.31	41.33	41.42
(d) values for Logistic ( $12 \times 12$ )						
	none $f(x)$ LOO			none $f(x)$ LOO		
	regular			antithetic		
MC	7447.38	22.83	22.86	45.62	45.61	45.75
QMC (lat.)	<b>22.56</b>	<b>22.56</b>	<b>22.61</b>	45.01	44.99	45.07
RQMC (l.)	<b>22.66</b>	<b>22.65</b>	<b>22.68</b>	45.30	45.32	45.41
(f) values for Gumbel ( $12 \times 12$ )						
	none $f(x)$ LOO			none $f(x)$ LOO		
	regular			antithetic		
MC	11642.74	52.89	46.11	—	—	—
QMC (lat.)	46.88	<b>45.41</b>	<b>45.48</b>	—	—	—
RQMC (l.)	58.12	<b>45.74</b>	<b>45.80</b>	—	—	—
(g) values for Cauchy ( $8 \times 8$ )						
	none	$f(x)$	LOO	none	$f(x)$	LOO
	regular			antithetic		
MC	249027.67	263426.66	255440.59	507004.19	525973.88	509764.25
QMC (lat.)	<b>2533.24</b>	<b>2532.93</b>	<b>2537.32</b>	<b>2531.24</b>	<b>2532.92</b>	<b>2537.35</b>
RQMC (l.)	251018.28	267124.91	264146.84	476293.00	507766.00	529030.06
(h) values for Cauchy ( $12 \times 12$ )						
	none	$f(x)$	LOO	none	$f(x)$	LOO
	regular			antithetic		
MC	1316801.88	1284078.38	1297748.25	2657888.00	2631427.25	2633413.50
QMC (lat.)	<b>12922.79</b>	<b>12922.31</b>	<b>12948.75</b>	<b>12931.28</b>	<b>12928.22</b>	<b>12945.27</b>
RQMC (l.)	1318297.38	1299869.75	1365709.75	2606723.50	2615697.50	2529304.00
(i) values for Laplace ( $8 \times 8$ )						
	none $f(x)$ LOO			none $f(x)$ LOO		
	regular			antithetic		
MC	2641.38	8.15	8.15	16.28	16.27	16.29
QMC (lat.)	<b>8.04</b>	<b>8.05</b>	<b>8.06</b>	16.01	16.00	16.04
RQMC (l.)	<b>8.09</b>	<b>8.09</b>	<b>8.10</b>	16.19	16.17	16.22
(j) values for Laplace ( $12 \times 12$ )						
	none $f(x)$ LOO			none $f(x)$ LOO		
	regular			antithetic		
MC	13593.82	<b>41.40</b>	<b>41.45</b>	82.73	82.71	82.92
QMC (lat.)	<b>41.06</b>	<b>41.07</b>	<b>41.16</b>	81.78	81.75	81.92
RQMC (l.)	<b>41.32</b>	<b>41.31</b>	<b>41.36</b>	82.62	82.64	82.80
(k) values for Triangular ( $8 \times 8$ )						
	none $f(x)$ LOO			none $f(x)$ LOO		
	regular			antithetic		
MC	3090.80	10.21	10.11	20.27	20.43	20.07
QMC (lat.)	<b>5.57</b>	<b>5.57</b>	<b>5.57</b>	10.17	10.18	10.20
RQMC (l.)	884.22	9.88	9.82	19.14	19.71	19.76
(l) values for Triangular ( $12 \times 12$ )						
	none $f(x)$ LOO			none $f(x)$ LOO		
	regular			antithetic		
MC	15975.60	49.73	49.89	99.81	99.32	100.31
QMC (lat.)	<b>28.28</b>	<b>28.28</b>	<b>28.34</b>	51.79	51.79	51.86
RQMC (l.)	4606.71	49.01	49.47	98.56	98.66	98.01