OPT2023: 15th Annual Workshop on Optimization for Machine Learning

Level Set Teleportation: the Good, the Bad, and the Ugly

Aaron Mishkin AARONPMISHKIN@CS.STANFORD.EDU
Department of Computer Science, Stanford University

Alberto Bietti ALBERTO@BIETTI.ME
CCM, Flatiron Institute

Robert M. Gower GOWERROBERT@GMAIL.COM
CCM, Flatiron Institute

Abstract

We study level set teleportation, a sub-routine which seeks to accelerate gradient methods by maximizing the gradient over the set of parameters with the same objective value. Since the descent lemma implies that gradient descent (GD) decreases the objective proportional to the gradient-norm, level-set teleportation maximizes guaranteed one-step progress. We prove level-set teleportation neither improves nor worsens the convergence of GD for strongly convex functions, while for convex functions teleportation can arbitrarily increase the distance to the global minima. To solve teleportation problems, we develop a projected-gradient-type method requiring only Hessian-vector products; we use our method to show that initializing GD with teleportation slightly under-performs standard initializations for both convex and non-convex optimization problems. As a result, we report a mixed picture: teleportation can be efficiently evaluated, but appears to offer marginal gains.

1. Introduction

We consider the minimization of a continuous, potentially non-convex function f. When the gradient of f is L-Lipschitz (i.e. f is L-smooth), the *descent lemma* (Bertsekas, 1997) implies gradient descent (GD) with step-size $\eta_k < 2/L$ makes progress proportional to the gradient-norm,

$$f(w_{k+1}) \leq f(w_k) - \eta_k \left(1 - \frac{\eta_k L}{2}\right) \|\nabla f(w_k)\|_2^2. \tag{1}$$

If all other quantities are held constant, then increasing the norm of the gradient increases the one-step progress guaranteed by smoothness. This implies gradient descent trajectories which maximize the observed gradients may converge faster than their naive counterparts. This same argument has been used to select new edges for incrementally growing a neural network (Evci et al., 2022).

Level set teleportation attempts to leverage the descent lemma by maximizing the gradient norm without changing the objective value. At an iteration k satisfying a pre-determined scheduling rule, level set teleportation solves the non-concave maximization problem,

$$w_k^+ \in \arg\max_w \frac{1}{2}\|\nabla f(w)\|_2^2 \quad \text{s.t.} \quad f(w) = f(w_k), \tag{2}$$

where the level set is $\mathcal{L}_k := \{w : f(w) = f(w_k)\}$. Zhao, Dehmamy, et al. (2023) show the Newton and gradient directions coincide after teleportation, meaning the next gradient step is equivalent to a

Newton update (see Fig. 1). For some functions, including quadratics, the gradient norm is also maximized everywhere along the gradient flow from w_k^+ (Zhao, Dehmamy, et al., 2023; Zhao, Gower, et al., 2023). This suggests that level set teleportation may be an effective heuristic for improving the convergence rate of gradient methods, particularly when used to initialize optimization.

A major barrier to evaluating the effectiveness of teleportation is the difficulty of solving Eq. (2). Previous work has instead focused on *symmetry teleportation* (Zhao, Dehmamy, et al., 2023; Zhao, Gower, et al., 2023), which restricts optimization to group symmetries of the objective. For example, neural networks with positively homogeneous activation functions are invariant under certain positive rescalings. Although group symmetries do not fully capture $\mathcal{L}_k$, Zhao, Dehmamy, et al. (2023) try to approximate level set teleportation by optimizing over parameterized group operators.

In contrast, we provide a simple algorithm for solving the level set teleportation problem based on sequential quadratic programming (SQP). Our method, which requires only Hessian-vector products, works by linearizing the level set constraint $f(w) = f(w_k)$ at each iteration and resembles a step of projected GD. The procedure is parameter-free—the step-size is selected automatically using a merit function—and convergence can be guaranteed using connections to existing SQP methods.

We use our algorithm to evaluate the effectiveness of level set teleportation for initializing optimization by computing w_0^+ and we also perform a limited theoretical investigation. Our contributions provide a mixed perspective on level set teleportation which we summarize as follows:

- **The Good**: We prove teleportation does not harm optimization when f is smooth and strongly-convex, although it cannot improve convergence in the worst case. We also provide a fast, parameter-free algorithm for solving teleportation and scale it to MNIST (LeCun et al., 1998).
- **The Bad**: We construct an example showing that level set teleportation can move arbitrarily far from the minimizers for non-strongly convex functions, making convergence guarantees unlikely in this setting. For deterministic updates, our experiments show initializing by teleportation only accelerates optimization initially and can lead to slow convergence later.
- **The Ugly**: Level set teleportation can improve convergence for specific problems, particularly over the first few epochs, yet appears to have little effect in the stochastic setting. As a result, it is difficult to strongly advocate for teleportation or to completely dismiss it outright.

1.1. Related Work

Level Set Teleportation: Armenta and Jodoin (2021) and Armenta, Judge, et al. (2020) use group symmetries to randomly perturb the weights of neural networks during training. Zhao, Dehmamy, et al. (2023) optimize over parametric symmetries, while Zhao, Ganev, et al. (2023) propose more sophisticated mappings, including data-dependent symmetry operators. Zhao, Gower, et al. (2023) give stronger guarantees on the gradient norm for symmetry teleportation on non-convex functions.

Symmetries in Optimization: Teleportation is closely connected to the notion of sharp minima in deep learning (Dinh et al., 2017; Hochreiter and Schmidhuber, 1997; Keskar et al., 2017). In particular, sharpness aware minimization (Foret et al., 2021) biases optimization towards "flat" regions with low curvature, while teleportation seeks large gradients to accelerate training. Neyshabur et al. (2015) propose Path-SGD, a gradient method which is invariant to rescaling symmetries, while Bamler and Mandt (2018) separate optimization into directions with symmetries and those without.

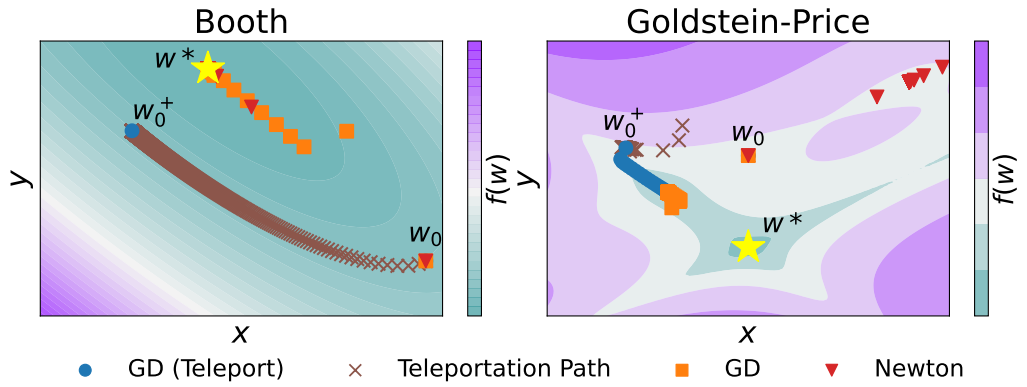

Figure 1: Initializing by level set teleportation on two test functions. Booth is quadratic and teleporting aligns w_0^+ with the maximum eigenvalue-eigenvector pair. Goldstein-Price is non-convex and teleporting pushes w_0^+ up a narrow "valley" from which convergence is slow.

2. Level Set Teleportation

We first analyze the effects of level set teleportation on the convergence of gradient methods. We assume that f is L-smooth, has at least one minimizer w^*, and is coercive. Coercivity implies the level sets are compact and guarantees the teleportation problem admits a finite solution.

Instead of solving Eq. (2), we focus on the more general sub-level set teleportation problem,

$$w_k^+ = \arg\max_w \frac{1}{2}\|\nabla f(w)\|_2^2 \quad \text{s.t.} \quad f(w) \leq f(w_k), \tag{3}$$

where the feasible set is the sub-level set $\mathcal{S}_k = \{w : f(w) \leq f(w_k)\}$. For convex f, Eq. (3) admits at least one solution on the boundary of $\mathcal{S}_k$; if f is strictly convex, then every solution is on the boundary and the relaxation is equivalent to level set teleportation (Lemma 5). However, sub-level set teleportation is acceptable even for non-convex functions since our overall goal is minimization.

Let $\mathcal{T} \subseteq \mathbb{N}$ be a teleportation schedule, meaning $w_{k+1} = w_k - \eta_k \nabla f(w_k^+)$ if $k \in \mathcal{T}$ and $w_{k+1} = w_k - \eta_k \nabla f(w_k)$ otherwise. Recall that f is μ-strongly convex if for every $w, w' \in \mathbb{R}^d$ it holds that

$$f(w) \geq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{\mu}{2}\|w - w'\|_2^2. \tag{4}$$

Our first result shows GD has the same convergence rate with and without teleportation.

Proposition 1 *Suppose f is L-smooth and μ strongly convex. Then gradient descent with step-size sequence $\{\eta_k\}$ and teleportation schedule $\mathcal{T}$ converges as*

$$f(w_k) - f(w^*) \leq \frac{L}{\mu} \prod_{i=0}^{k} \left[\max\left\{(1 - \eta_i L)^2, (1 - \eta_i \mu)^2\right\}\right] (f(w_0) - f(w^*)).$$

Moreover, this rate is tight in the worst case.

See Appendix A for proof. While Zhao, Gower, et al. (2023) give an upper bound for symmetry teleportation with GD under the weaker Polyak-Łojasiewicz condition (Karimi et al., 2016), the

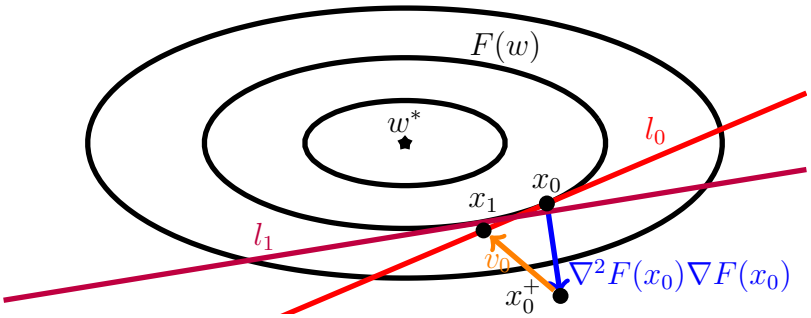

Figure 2: Illustration of our method for solving level set teleportation on a convex quadratic. The algorithm combines gradient-ascent with projections onto a linearized level-set.

final rate is worse and they do not study lower bounds. Next we consider non-strongly convex f and prove sub-level set teleportation can arbitrarily increase the distance to the minimizer.

Proposition 2 *For every $C > 0$, there exists smooth, convex f such that teleportation satisfies*

$$\|w_k^+ - w^*\|_2 \geq C\|w_k - w^*\|_2.$$

Although Proposition 2 is not a non-convergence result for GD with teleportation, it implies standard proofs for smooth, convex functions which start by expanding $\|w_{k+1} - w^*\|_2^2$ are likely to fail.

3. Evaluating the Teleportation Operator

Sub-level set teleportation requires solving a general non-linear programming problem. Although we could apply standard SQP methods, computing the Hessian of the teleportation objective requires third-order derivatives of f, which is not feasible for large-scale problems. Instead, we develop an iterative projected-gradient-type method which is scalable and requires only Hessian-vector products. We denote by x_t the iterates of our method for solving teleportation, with $x_0 = w_k$.

For general f, the sub-level set $\mathcal{S}_k$ is non-convex and does not admit an efficient projection operator. In contrast, the linearization of this constraint around an iterate t x_t yields a single half-space,

$$\tilde{\mathcal{S}}_k(x_t) := \{w : l_k(x) := f(x_t) + \langle \nabla f(x_t), w - x_t \rangle \leq f(w_k)\},$$

for which projections are easy. To obtain a tractable algorithm, we consider maximizing a penalized linearization of the log-gradient-norm subject to this constraint,

$$x_{t+1} = \underset{x \in \tilde{\mathcal{S}}_k(x_t)}{\arg\max} \left\{ \frac{1}{2}\log(\|\nabla f(x_t)\|_2^2) + \left\langle \frac{\nabla^2 f(x_t)\nabla f(x_t)}{\|\nabla f(x_t)\|_2^2}, x - x_t \right\rangle - \frac{1}{\rho_t}\|x - x_t\|_2^2 \right\}. \tag{5}$$

Taking the logarithm of the objective implicitly encodes positivity of the gradient norm and leads to a normalized update rule, as we show next.

Proposition 3 *The solution to Eq. (5) is given by*

$$\begin{aligned} v_t &= -\left(\rho_t \left\langle \nabla f(x_t), \nabla^2 f(x_t)\nabla f(x_t) \right\rangle + f(x_t) - f(w_k)\right)_+ \nabla f(x_t), \\ x_{t+1} &= x_t + \left(\rho_t \nabla^2 f(x_t)\nabla f(x_t) + v_t\right) / \|\nabla f(x_t)\|_2^2. \end{aligned} \tag{6}$$

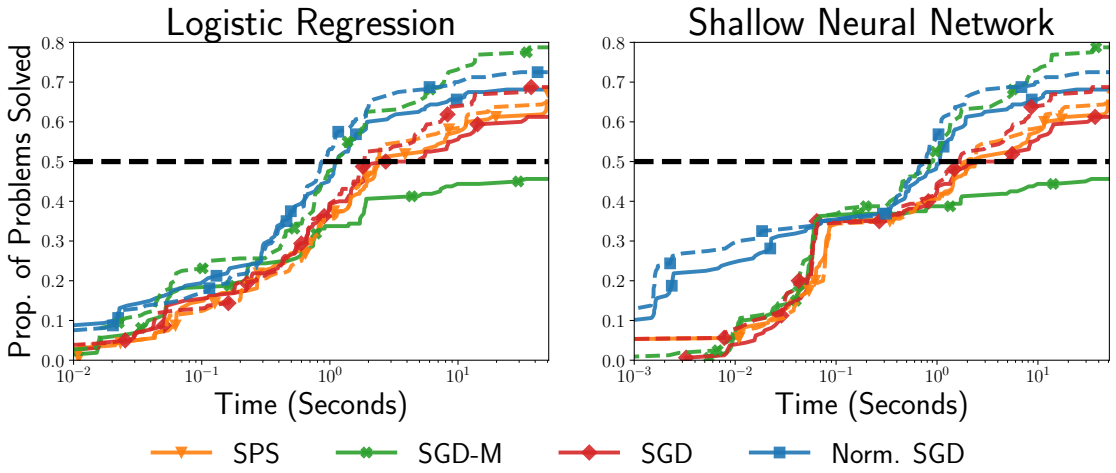

Figure 3: Performance profile comparing optimization with (solid lines) and without (dashed lines) initialization by teleportation. A problem is solved when $(f(w_k) - f(w^*))/f(w^*) \leq 0.5$, where $f(w^*)$ is the smallest objective found by any method. Performance is judged by comparing time to a fixed proportion of problems solved (see dashed line at 50%).

This iteration is equivalent to a step of projected gradient descent with a linearized sub-level set constraint; see Fig. 2 for an illustration. It is also a step of sequential quadratic programming using the crude estimator $\mathbf{I}/\rho_t \approx \nabla^2 f(x_t)$. Torrisi et al. (2018) leverage this fact to show that projected-gradient algorithms with linearized constraints are convergent, although they require an additional relaxation step $\hat{x}_{k+1} = \alpha x_{t+1} + (1-\alpha)x_t$. We have not found this to be necessary in practice.

3.1. Computing the Step-size

A major disadvantage of our update is that it requires a step-size $\rho_t > 0$. Since teleportation is a sub-routine of a larger optimization procedure, tuning ρ_t for good performance is not acceptable. Following standard practice for SQP methods (see, e.g. Nocedal and Wright (1999, Theorem 18.2)), we instead select ρ_t using line-search on an Armijo-type condition (Armijo, 1966),

$$\phi_\gamma(x_{t+1}) \leq \phi_\gamma(x_t) + \frac{1}{2} D_\phi(x_t, x_{t+1} - x_t), \tag{7}$$

where $\phi_\gamma(x) = -\frac{1}{2}\|\nabla f(x)\|_2^2 + \gamma\left(f(x) - f(w_k)\right)_+$, $\gamma > 0$ controls the penalty strength, and $D_\phi(x_t, d)$ is the directional derivative of ϕ_γ. Setting γ sufficiently large gives a descent direction.

Proposition 4 *Let $q_t = \nabla^2 f(x_t)\nabla f(x_t)$ and suppose $\gamma_t > \frac{\langle q_t, v_t\rangle}{\|\nabla f(x_t)\|_2^2 (f(x_t) - f(w_k))}$. Then $x_{t+1} - x_t$ is a descent direction of ϕ_{γ_t}, and the line-search condition simplifies to*

$$\phi_{\gamma_t}(x_{t+1}) \leq -\frac{1}{2}\|\nabla f(x_t)\|_2^2 + \left(\langle q_t, v_t\rangle - \rho_t \|q_t\|_2^2\right)/\|\nabla f(x_t)\|_2^2. \tag{8}$$

When $v_t = 0$ the update reduces to gradient ascent and $\gamma_t = 0$ is immediately sufficient for progress. Proposition 4 provides a recipe for computing step-sizes using backtracking line-search. See Appendix B for details on termination criteria and a full description of our algorithm.

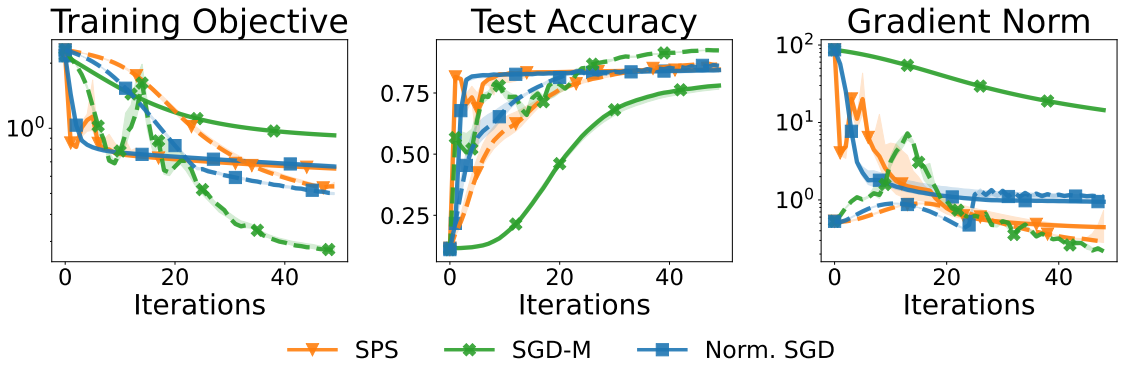

Figure 4: Performance of full-batch optimizers with (solid) and without (dashed) initialization by teleportation for training a ReLU MLP with two hidden layers of size 100 on MNIST.

4. Experiments

Solving the Teleportation Problem: Fig. 1 shows the convergence path of our teleportation solver on two test functions (Goldstein and Price, 1971). See Fig. 5 in Appendix C.1 for additional results showing our method converges to an approximate KKT point when teleporting on MNIST.

Initialization by Teleportation: Fig. 3 presents a performance profile (Dolan and Moré, 2002) comparing full-batch gradient descent (SGD), the Polyak step-size (SPS) (Loizou et al., 2021; Polyak, 1987), and normalized gradient descent (Norm. SGD) with (solid lines) and without (dashed lines) initialization by teleportation on 160 problems from the UCI repository (Asuncion and Newman, 2007). We find that teleportation does not improve the performance of gradient methods in general. Out of all three methods, only normalized gradient descent is competitive when used with teleportation and it only outperforms the standard initialization early in optimization. See Fig. 7 for similar results in the stochastic setting and Fig. 10 for special cases where teleportation is advantageous.

Image Classification: We perform additional experiments with ReLU networks on MNIST (Fig. 4) to confirm our observations. We find that teleportation significantly increases the gradient norm along the optimization path, but methods initialized by teleporting require smaller step-sizes. As a result, the intuition from the descent lemma fails and teleportation stalls. See Appendix C.1 for additional results on Fashion MNIST (Xiao et al., 2017) and extensions to the stochastic setting.

5. Conclusion

Despite recent work advocating for level set teleportation as an optimization sub-routine for gradient methods, little work has been done to solve teleportation problems or evaluate their practical utility. We rectify this and study (sub)-level set teleportation in detail; we prove new theoretical guarantees for gradient descent with teleportation, derive a novel algorithm for solving teleportation problems, and evaluate the performance of teleportation on a large suite of problems. Our results reveal the surprisingly mixed performance of teleportation in both theory and practice and we advocate a balanced viewpoint that includes all its aspects — good, bad, and ugly.

References

Armenta, Marco and Pierre-Marc Jodoin (2021). "The representation theory of neural networks". In: *Mathematics* 9.24, p. 3216.

Armenta, Marco, Thierry Judge, et al. (2020). "Neural Teleportation". In: *CoRR* abs/2012.01118.

Armijo, Larry (1966). "Minimization of functions having Lipschitz continuous first partial derivatives". In: *Pacific Journal of mathematics* 16.1, pp. 1–3.

Asuncion, Arthur and David Newman (2007). *UCI machine learning repository*.

Bamler, Robert and Stephan Mandt (2018). "Improving optimization for models with continuous symmetry breaking". In: *International Conference on Machine Learning*. PMLR, pp. 423–432.

Bertsekas, Dimitri P (1997). "Nonlinear programming". In: *Journal of the Operational Research Society* 48.3, pp. 334–334.

Dinh, Laurent et al. (2017). "Sharp Minima Can Generalize For Deep Nets". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Vol. 70. Proceedings of Machine Learning Research, pp. 1019–1028.

Dolan, Elizabeth D and Jorge J Moré (2002). "Benchmarking optimization software with performance profiles". In: *Mathematical programming* 91, pp. 201–213.

Evci, Utku et al. (2022). "GradMax: Growing Neural Networks using Gradient Information". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=qjN4h_wwUO.

Fernández-Delgado, Manuel et al. (2014). "Do we need hundreds of classifiers to solve real world classification problems?" In: *The journal of machine learning research* 15.1, pp. 3133–3181.

Foret, Pierre et al. (2021). "Sharpness-aware Minimization for Efficiently Improving Generalization". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Goldstein, Allen A. and J. F. Price (1971). "On descent from local minima". In: *Mathematics of Computation* 25, pp. 569–574.

He, Kaiming et al. (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1026–1034.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Flat Minima". In: *Neural Comput.* 9.1, pp. 1–42.

Karimi, Hamed, Julie Nutini, and Mark Schmidt (2016). "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition". In: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*. Vol. 9851. Lecture Notes in Computer Science, pp. 795–811.

Keskar, Nitish Shirish et al. (2017). "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

LeCun, Yann et al. (1998). "Gradient-based learning applied to document recognition". In: *Proc. IEEE* 86.11, pp. 2278–2324.

Loizou, Nicolas et al. (2021). "Stochastic Polyak Step-size for SGD: An Adaptive Learning Rate for Fast Convergence". In: *The 24th International Conference on Artificial Intelligence and Statistics,*

AISTATS 2021, April 13-15, 2021, Virtual Event. Vol. 130. Proceedings of Machine Learning Research, pp. 1306–1314.

Nesterov, Yurii E. (2004). *Introductory Lectures on Convex Optimization - A Basic Course*. Vol. 87. Applied Optimization. Springer.

Neyshabur, Behnam, Ruslan Salakhutdinov, and Nathan Srebro (2015). "Path-SGD: Path-Normalized Optimization in Deep Neural Networks". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2422–2430.

Nocedal, Jorge and Stephen J Wright (1999). *Numerical optimization*. Springer.

Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035.

Polyak, Boris T (1987). "Introduction to optimization". In.

Torrisi, Giampaolo et al. (2018). "A Projected Gradient and Constraint Linearization Method for Nonlinear Model Predictive Control". In: *SIAM J. Control. Optim.* 56.3, pp. 1968–1999.

Xiao, Han, Kashif Rasul, and Roland Vollgraf (2017). "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms". In: *arXiv preprint arXiv:1708.07747*.

Zhao, Bo, Nima Dehmamy, et al. (2023). *Symmetry Teleportation for Accelerated Optimization*. arXiv: 2205.10637 [cs.LG].

Zhao, Bo, Iordan Ganev, et al. (2023). "Symmetries, Flat Minima, and the Conserved Quantities of Gradient Flow". In: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Zhao, Bo, Robert M. Gower, et al. (2023). "Improving Convergence and Generalization Using Parameter Symmetries". In: *CoRR* abs/2305.13404.

Appendix A. Proofs

Lemma 5 *Suppose f is (strictly) convex. Then at least one (every) solution to the sub-level set teleportation problem (Eq. (3)) is a solution to the level set teleportation problem (2).*

Proof Let $w(t) = w + t\nabla f(w)$. From convexity,

$$f(w(t)) - f(w) \geq t\|\nabla f(w)\|_2^2$$
$$f(w) - f(w(t)) \geq t\left\langle \nabla f(w(t)), \nabla f(w)\right\rangle_2^2.$$

Adding these inequalities and using Cauchy-Schwarz,

$$\implies \|\nabla f(w(t))\|_2 \geq \|\nabla f(w)\|_2.$$

That is, the gradient norm is monotone non-decreasing when f is convex. At least one solution to the maximization problem must occur on the boundary of the sub-level set, which completes the first part of the proof. For the second, simply note that the inequalities hold strictly if f is strictly convex, implying that every solution must be on the boundary. ■

Proposition 1 *Suppose f is L-smooth and μ strongly convex. Then gradient descent with step-size sequence $\{\eta_k\}$ and teleportation schedule $\mathcal{T}$ converges as*

$$f(w_k) - f(w^*) \leq \frac{L}{\mu}\prod_{i=0}^{k}\left[\max\left\{(1-\eta_i L)^2, (1-\eta_i\mu)^2\right\}\right](f(w_0) - f(w^*)).$$

Moreover, this rate is tight in the worst case.

Proof First we show the upper bound. Since f is L-smooth and μ strongly convex, ∇f satisfies the following inequality:

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{\mu L}{\mu + L}\|x-y\|_2^2 + \frac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|. \tag{9}$$

This is sometimes called coercivity of the gradient; see, for example, Nesterov (2004, Theorem 2.1.12). Suppose $k \in \mathcal{T}$. Then,

$$\begin{aligned}
\|w_{k+1} - w^*\|_2^2 &= \|w_k^+ - \eta_k\nabla f(w_k^+) - w^*\|_2^2 \\
&= \|w_k^+ - w^*\|_2^2 - 2\eta_k\left\langle \nabla f(w_k^+), w_k^+ - w^*\right\rangle + \eta_k^2\|\nabla f(w_k^+)\|_2^2 \\
&\leq \|w_k^+ - w^*\|_2^2 - 2\eta_k\left(\frac{\mu L}{\mu+L}\|w_k^+ - w^*\|_2^2 + \frac{1}{L+\mu}\|\nabla f(w_k^+)\|\right) \\
&\qquad + \eta_k^2\|\nabla f(w_k^+)\|_2^2 \\
&= \left(1 - \frac{2\eta_k\mu L}{\mu+L}\right)\|w_k^+ - w^*\|_2^2 + \eta_k\left(\eta_k - \frac{2}{\mu+L}\right)\|\nabla f(w_k^+)\|_2^2 \\
&\leq \left(1 - \frac{2\eta_k\mu L}{\mu+L}\right)\|w_k^+ - w^*\|_2^2 \\
&\qquad + \eta_k\max\left\{L^2\left(\eta_k - \frac{2}{\mu+L}\right), \mu^2\left(\eta_k - \frac{2}{\mu+L}\right)\right\}\|w_k^+ - w^*\|_2^2 \\
&= \max\left\{(1-\eta_k L)^2, (1-\eta_k\mu)^2\right\}\|w_k^+ - w^*\|_2^2,
\end{aligned}$$

where we have used μ strong convexity and L-smoothness to bound the gradient norm depending on the step-size η_k. Now we convert from iterates to function values using smoothness and strong convexity again to obtain

$$\begin{aligned}\implies f(w_{k+1}) - w^* &\leq \frac{L}{\mu}\max\left\{(1-\eta_k L)^2, (1-\eta_k\mu)^2\right\}\left(f(w_k^+) - f(w^*)\right) \\ &\leq \frac{L}{\mu}\max\left\{(1-\eta_k L)^2, (1-\eta_k\mu)^2\right\}\left(f(w_k) - f(w^*)\right),\end{aligned}$$

where the last inequality follows from the definition of sub-level set teleportation. If $k \notin \mathcal{T}$, then the proof proceeds identically without needing this final step. In either case, recursing on this expression is now sufficient to give the final result.

For the lower bound, suppose $f(w) = \frac{1}{2}\|w\|_2^2$. Clearly f is 1-smooth and strongly convex with $\mu = 1$. The unique minimizer is simply $w^* = 0$. Starting from an arbitrary w_0, each iteration of GD with teleportation has the following recursion:

$$w_{k+1} = w_k^+ - \eta_k w_k^+ = (1-\eta_k)w_k^+,$$

and the objective evolves trivially as

$$\begin{aligned}\frac{1}{2}\|w_{k+1}\|_2^2 &= (1-\eta_k)^2\frac{1}{2}\|w_k^+\|_2^2 \\ &= (1-\eta_k)^2\frac{1}{2}\|w_k\|_2^2,\end{aligned}$$

where the second equality uses Lemma 5 to guarantee that the solution to sub-level set teleportation lies on the level set $\mathcal{L}_k$ and the fact that every point $x \in \mathcal{L}_k$ satisfies $\|x\|_2 = \|w_k\|$. Thus, each step of gradient descent makes progress exactly matching the convergence rate for general smooth, strongly convex functions regardless of teleportation. This completes the lower-bound. ■

Proposition 2 *For every $C > 0$, there exists smooth, convex f such that teleportation satisfies*

$$\|w_k^+ - w^*\|_2 \geq C\|w_k - w^*\|_2.$$

Proof We assume for convenience that $w = (x, y)$, $x > 1$ and $y > 0$. These assumptions can be relaxed by modifying our construction, but the calculations are tedious. Let $\epsilon > 0$ such that $\alpha, y > \epsilon$ and $x + \epsilon y > \alpha\epsilon + 1$ hold, where both conditions can be satisfied by taking ϵ sufficiently small.

Let g_δ be the Huber function defined by

$$g_\delta(x) = \begin{cases}\frac{1}{2}x^2 & \text{if } x \leq \delta \\ \delta\left(|x| - \delta/2\right) & \text{otherwise.}\end{cases} \tag{10}$$

and consider the objective function

$$f_{(\epsilon,\alpha)}(x, y) = g_1(x) + g_\epsilon(y) + \frac{1}{2}\mathbb{1}_{y\geq\alpha}\left(y - \alpha\right)^2. \tag{11}$$

We first show that there exists $x' = 1$ and some $y' > \alpha$ satisfying

$$f_{(\epsilon,\alpha)}(x', y') = f_{(\epsilon,\alpha)}(x, y).$$

For this to hold, we must have

$$\begin{aligned} & f_{(\epsilon,\alpha)}(x', y') = 1 + \epsilon y' - \frac{\epsilon}{2} + \frac{1}{2}\left(y' - \alpha\right)^2 = x + \epsilon y - \frac{\epsilon}{2} = f_{(\epsilon,\alpha)}(x, y) \\ \iff & (y')^2 + 2\left(\epsilon - \alpha\right) y' + \left(\alpha^2 - 2x - 2\epsilon y + 2\right) = 0 \\ \iff & y' = \alpha - \epsilon \pm \sqrt{(\epsilon - \alpha)^2 - \alpha^2 + 2x + 2\epsilon y - 2}. \end{aligned}$$

In particular, for $y' > \alpha$, it must hold that

$$\begin{aligned} & (\epsilon - \alpha)^2 - \alpha^2 + 2x + 2\epsilon y - 2 > \epsilon^2 \\ \iff & x + \epsilon y > \alpha\epsilon + 1, \end{aligned}$$

where this last condition is guaranteed by assumption on ϵ. We conclude that (x', y') is on the level set as desired.

The gradient of $f_{(\epsilon,\alpha)}$ is easy to calculate as

$$\nabla f_{(\epsilon,\alpha)}(x, y) = \begin{cases} \text{sign}(x) + \epsilon \cdot \text{sign}(y) + \mathbb{1}_{y \geq \alpha}\left(y - \alpha\right) & \text{if } |x| \geq 1, |y| \geq \epsilon \\ x + \epsilon \cdot \text{sign}(y) + \mathbb{1}_{y \geq \alpha}\left(y - \alpha\right) & \text{if } |x| \leq 1, |y| \geq \epsilon \\ \text{sign}(x) + \epsilon y + \mathbb{1}_{y \geq \alpha}\left(y - \alpha\right) & \text{if } |x| \geq 1, |y| \leq \epsilon \\ x + \epsilon y + \mathbb{1}_{y \geq \alpha}\left(y - \alpha\right) & \text{if } |x| \leq 1, |y| \leq \epsilon. \end{cases}$$

In particular, the gradient norm at (x', y') is given by

$$\|\nabla f_{(\epsilon,\alpha)}(x', y')\|_2^2 = 1 + \epsilon^2 + (y' - \alpha)^2.$$

A straightforward case analysis reveals that for every $\bar{x}$, $\bar{y}$ such that $\bar{y} < \alpha$,

$$\|\nabla f_{(\epsilon,\alpha)}(\bar{x}, \bar{y})\|_2^2 < \|\nabla f_{(\epsilon,\alpha)}(x', y')\|_2^2.$$

That is, the maximizer of the gradient norm on the level set must satisfy $y^+ \geq \alpha$. We conclude that

$$\|w^+ - w^*\|_2^2 = \|(x^+, y^+)\|_2^2 \geq \alpha^2,$$

and, choosing $\alpha = C \cdot \|(x, y)\|_2$,

$$\|w^+ - w^*\|_2^2 \geq C\|w - w^*\|_2^2.$$

■

Proposition 3 *The solution to Eq. (5) is given by*

$$\begin{aligned} v_t &= -\left(\rho_t \left\langle \nabla f(x_t), \nabla^2 f(x_t) \nabla f(x_t) \right\rangle + f(x_t) - f(w_k)\right)_+ \nabla f(x_t), \\ x_{t+1} &= x_t + \left(\rho_t \nabla^2 f(x_t) \nabla f(x_t) + v_t\right) / \|\nabla f(x_t)\|_2^2. \end{aligned} \tag{6}$$

Proof We proceed by case analysis. Let

$$\bar{x} = \arg\max_{x \in \mathbb{R}^d} \left\{ \frac{1}{2} \log(\|\nabla f(x_t)\|_2^2) + \left\langle \frac{\nabla^2 f(x_t) \nabla f(x_t)}{\|\nabla f(x_t)\|_2^2}, x - x_t \right\rangle - \frac{1}{\rho_t} \|x - x_t\|_2^2 \right\}.$$

Case 1: $\bar{x} \in \tilde{\mathcal{S}}_k(x_t)$. Then $\bar{x}$ satisfies the linearized constraint and $x_{t+1} = \bar{x}$ must hold. It is straightforward to compute that

$$x_{t+1} = x_t - \rho \nabla^2 f(x_t) \nabla f(x_t).$$

Substituting this value into the linearization of the half-space and using $x_{t+1} \in \tilde{\mathcal{S}}_k(x_t)$, we find

$$\rho \nabla f(x_t)^\top \nabla^2 f(x_t) \nabla f(x_t) + f(x_t) - f(w_k) \leq 0.$$

Case 2: $\bar{x} \notin \tilde{\mathcal{S}}_k(x_t)$. Then the solution lies on the boundary of the half-space constraint and is given by projecting $\bar{x}$ onto

$$\tilde{\mathcal{L}}_k(x_t) = \{x : f(x_t) + \langle \nabla f(x_t), x - x_t \rangle = f(w_k)\}.$$

The Lagrangian of this problem is

$$L(x, \lambda) = \frac{1}{2} \|x - \bar{x}\|_2^2 + \lambda \left(f(x_t) + \langle \nabla f(x_t), x - x_t \rangle - f(w_k) \right).$$

Minimizing over x yields

$$x_{t+1} = \bar{x} - \lambda \nabla f(x_t),$$

which shows that the dual function is given by,

$$d(\lambda) = -\frac{1}{2} \lambda^2 \|\nabla f(x_t)\|_2^2 + \lambda \left(f(x_t) + \langle \nabla f(x_t), \bar{x} - x_t \rangle - f(w_k) \right).$$

This is a concave quadratic and maximizing over λ gives the following dual solution:

$$\lambda^* = \frac{f(x_t) - f(w_k) + \rho \left\langle \nabla f(x_t), \nabla^2 f(x_t) \nabla f(x_t) \right\rangle}{\|\nabla f(x_t)\|_2^2},$$

where have expanded the value of $\bar{x}$. Plugging this value back into the expression for x_{t+1},

$$\begin{aligned} x_{t+1} = x_t &- \rho \nabla^2 f(x_t) \nabla f(x_t) \\ &- \frac{\rho \left\langle \nabla f(x_t), \nabla^2 f(x_t) \nabla f(x_t) \right\rangle + f(x_t) - f(w_k)}{\|\nabla f(x_t)\|_2^2} \nabla f(x_t). \end{aligned}$$

This completes the second case. Putting the analysis together, we obtain the desired result:

$$\begin{aligned} x_{t+1} = x_t &- \rho \nabla^2 f(x_t) \nabla f(x_t) \\ &- \left(\frac{\rho \left\langle \nabla f(x_t), \nabla^2 f(x_t) \nabla f(x_t) \right\rangle + f(x_t) - f(w_k)}{\|\nabla f(x_t)\|_2^2} \right)_+ \nabla f(x_t). \end{aligned}$$

■

Lemma 6 *The directional derivative of the merit function satisfies*

$$D_\phi(x_t; d_t) \leq \frac{\left\langle \nabla^2 f(x_t)\nabla f(x_t), v_t \right\rangle - \rho_t \|\nabla^2 f(x_t)\nabla f(x)\|_2^2}{\|\nabla f(x_t)\|_2^2} - \gamma \left(f(x_t) - f(w_k)\right)_+ . \tag{12}$$

As a result, if

$$\gamma > \frac{\left\langle \nabla^2 f(x_t)\nabla f(x_t), v_t \right\rangle}{\|\nabla f(x_t)\|_2^2 (f(x_t) - f(w_k))},$$

then $x_{t+1} - x_t$ is a descent direction for ϕ_γ at x_t.

Proof Let $d_t = x_{t+1} - x_t$. Define $\Delta_t(\alpha) = \phi_\gamma(w^t + \alpha d_t) - \phi_\gamma(x_t)$. Using first-order Taylor expansions, we have

$$\begin{aligned}
\Delta_t(\alpha) &= -\frac{1}{2}\|\nabla f(x_t + \alpha d_t)\|_2^2 + \gamma \left(f(x_t + \alpha d_t) - f(w_k)\right)_+ + \frac{1}{2}\|\nabla f(x_t)\|_2^2 - \gamma \left(f(x_t) - f(w_k)\right)_+ \\
&= -\alpha \left\langle \nabla^2 f(x_t)\nabla f(x_t), d_t \right\rangle + \gamma \left(f(x_t) + \alpha \left\langle \nabla f(x_t), d_t \right\rangle - f(w_k)\right)_+ - \gamma \left(f(x_t) - f(w_k)\right)_+ \\
&\qquad + O(\alpha^2) \\
&\leq -\alpha \left\langle \nabla^2 f(x_t)\nabla f(x_t), d_t \right\rangle + \gamma(1-\alpha) \left(f(x_t) - f(w_k)\right)_+ - \gamma \left(f(x_t) - f(w_k)\right)_+ + O(\alpha^2),
\end{aligned}$$

where we have used $\langle \nabla f(x_t), d_t \rangle \leq f(w_k) - f(x_t)$ from the definition of d_t. Simplifying, we obtain,

$$\Delta_t(\alpha) \leq= -\alpha \left\langle \nabla^2 f(x_t)\nabla f(x_t), d_t \right\rangle - \gamma\alpha \left(f(x_t) - f(w_k)\right)_+ + O(\alpha^2).$$

Dividing both sides by α and taking the limit as $\alpha \to 0$ shows that

$$\begin{aligned}
D_\phi(x_t; d_t) &\leq= -\left\langle \nabla^2 f(x_t)\nabla f(x_t), d_t \right\rangle + \gamma \left(f(x_t) - f(w_k)\right)_+ \\
&= \left(\left\langle \nabla^2 f(x_t)\nabla f(x_t), v_t \right\rangle - \rho_t \|\nabla^2 f(x_t)\nabla f(x)\|_2^2\right) / \|\nabla f(x_t)\|_2^2 - \gamma \left(f(x_t) - f(w_k)\right)_+ .
\end{aligned}$$

■

Proposition 4 *Let $q_t = \nabla^2 f(x_t)\nabla f(x_t)$ and suppose $\gamma_t > \frac{\langle q_t, v_t \rangle}{\|\nabla f(x_t)\|_2^2 (f(x_t) - f(w_k))}$. Then $x_{t+1} - x_t$ is a descent direction of ϕ_{γ_t}, and the line-search condition simplifies to*

$$\phi_{\gamma_t}(x_{t+1}) \leq -\frac{1}{2}\|\nabla f(x_t)\|_2^2 + \left(\langle q_t, v_t \rangle - \rho_t \|q_t\|_2^2\right) / \|\nabla f(x_t)\|_2^2. \tag{8}$$

Proof The first part of the proof follows immediately from Lemma 6. Substituting the expression for the directional derivative into the line-search condition,

$$\begin{aligned}
\phi_\gamma(x_{t+1}) &\leq -\frac{1}{2}\|\nabla f(x_t)\|_2^2 + \gamma \left(f(x_t) - f(w_k)\right)_+ - \gamma \left(f(x_t) - f(w_k)\right)_+ \\
&\qquad + \left(\left\langle \nabla^2 f(x_t)\nabla f(x_t), v_t \right\rangle - \rho_t \|\nabla^2 f(x_t)\nabla f(x)\|_2^2\right) / \|\nabla f(x_t)\|_2^2 \\
&= -\frac{1}{2}\|\nabla f(x_t)\|_2^2 + \left(\left\langle \nabla^2 f(x_t)\nabla f(x_t), v_t \right\rangle - \rho_t \|\nabla^2 f(x_t)\nabla f(x)\|_2^2\right) / \|\nabla f(x_t)\|_2^2,
\end{aligned}$$

which is straightforward to check in practice. ■

Proposition 7 *Suppose g is a loss function and $f(w) = g(h_w(X), y)$, where*

$$h_w(X) = \phi(W_l\phi(\dots W_2(\phi(W_1X)))$$

is the prediction function of a neural network with weights $w = (W_l, \dots, W_1)$, $l \geq 2$. If the activation function ϕ is positively homogeneous such that $\lim_{\beta\to\infty} \phi(\beta) = \infty$ and $\lim_{\beta\to\infty} \phi'(\beta) < \infty$, then optimal value of the sub-level set teleportation problem is unbounded and Eq. (3) does not admit a finite solution.

Proof For simplicity, we prove the result in the scalar case for $l = 2$, although it immediately generalizes. Since ϕ is positively homogeneous, we have

$$w_2\phi(w_1x) = \alpha w_2\phi((w_1/\alpha)x).$$

Let $\tilde{w}_2 = \alpha w_2$ and $\tilde{w}_1 = w_1/\alpha$. Define $v = w_2\phi(w_1x) = \tilde{w}_2\phi(\tilde{w}_1x)$. Then gradients with respect to the first and second layers are given by

$$\begin{aligned}
\frac{\partial}{\partial\tilde{w}_1}f(w) &= \frac{\partial}{\partial v}g(v)w_2\phi'(\tilde{w}_1x)x \\
&= \alpha\frac{\partial}{\partial v}g(v)w_2\phi'(w_1x/\alpha)x \\
\frac{\partial}{\partial\tilde{w}_2}f(w) &= \frac{\partial}{\partial v}g(v)\phi(\tilde{w}_1x) \\
&= \frac{\partial}{\partial v}g(v)\phi(w_1x/\alpha).
\end{aligned}$$

Taking the limit as $\alpha \to 0$, we see that

$$\begin{aligned}
\frac{\partial}{\partial\tilde{w}_1}f(w) &\to 0 \\
\frac{\partial}{\partial\tilde{w}_2}f(w) &\to \infty,
\end{aligned}$$

by assumption on ϕ. Thus, there exists a diverging sequence of points on the level set whose gradient norm is also diverging. Since the level sets are unbounded, the objective is not coercive and the problem is ill-posed. This completes the proof. ■

Algorithm 1: Sub-level Set Teleportation

Input: Iterate: w_k; Initial Step-size: ρ; Tolerances: ϵ, δ.

$x_0 \leftarrow w_k$

$q_0 \leftarrow \nabla^2 f(x_0)\nabla f(x_0)$ // Update direction

while $\|\mathbf{P}_t q_t\|_2 > \epsilon$ *or* $f(x_t) - f(w_k) > \delta$ **do** // KKT conditions

  $v_t \leftarrow -\left(\rho \langle q_t, \nabla f(x_t)\rangle + f(x_t) - f(w_k)\right)_+ \nabla f(x_t)$ // Correction factor

  $x_{t+1} \leftarrow x_t + (\rho \cdot q_t + v_t)/\|\nabla f(x_t)\|_2^2$

  while $\phi_{\gamma_t}(x_{t+1}) > \frac{1}{2}\|\nabla f(x_t)\|_2^2 + (\langle q_t, v_t\rangle - \rho\|q_t\|_2^2)/\|\nabla f(x_t)\|_2^2$ **do** // Line-search

    $\rho \leftarrow \rho/2$

    $v_t \leftarrow -\left(\rho \langle q_t, \nabla f(x_t)\rangle + f(x_t) - f(w_k)\right)_+ \nabla f(x_t)$

    $x_{t+1} \leftarrow x_t + (\rho \cdot q_t + v_t)/\|\nabla f(x_t)\|_2^2$

  end

  $q_{t+1} \leftarrow \nabla^2 f(x_t)\nabla f(x_t)$

end

Output: x_{t+1}

Appendix B. Additional Algorithmic Details

Now we briefly discuss how to terminate our algorithm. Suppose that $\bar{x}$ is a local maximum of Eq. (3). If the linear independence constraint qualification (LICQ) holds, then $\bar{x}$ must satisfy the KKT conditions for some $\lambda > 0$ (Bertsekas, 1997),

$$\begin{aligned} \nabla^2 f(x_t)\nabla f(x_t)/\|\nabla f(x_t)\| + \lambda \nabla f(x_t) = 0, \\ f(x_t) \leq f(w_k) \quad \text{and} \quad \lambda(f(x_t) - f(w_k)) = 0. \end{aligned} \tag{13}$$

That is, $\nabla^2 f(x_t)\nabla f(x_t) \propto \nabla f(x_t)$ and the sub-level set constraint is satisfied. The teleportation problem satisfies LICQ unless $\nabla f(\bar{x}) = 0$; in this case, stationarity at $\bar{x}$ implies stationarity over all of $\mathcal{S}_k$ and teleporting is not interesting. Thus, we assume LICQ holds and combine termination based on the KKT conditions with line-search to give a complete solver in Algorithm 1.

Appendix C. Experiments

C.1. Additional Experiments

Now we give experimental results which could not be included in the main paper due to space constraints.

Scaling Teleportation to MNIST: To further demonstrate the effectiveness of teleportation algorithm, we solve sub-level set teleportation for a two-layer MLP with fifty hidden units and soft-plus activations on the MNIST dataset. We use weight-decay regularization to ensure the objective is coercive. Fig. 5 shows the norm of the network gradient (i.e. the teleportation objective), the KKT residual (13), and violation of the constraints during teleportation. Our algorithm converges to a KKT point where the gradient norm is two orders of magnitude larger than that at the standard Kaiming initialization (He et al., 2015).

Effects of Regularization: The teleportation problem for ReLU networks does not admit a finite solution without additional regularization (Proposition 7). Fig. 6 confirms this result and shows

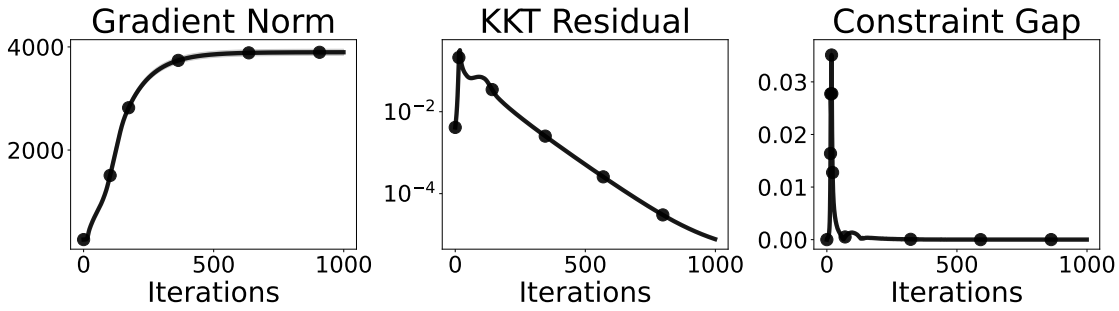

Figure 5: Sub-level set teleportation for a two-layer MLP with 50 hidden units on MNIST. Our algorithm finds an approximate KKT point despite the non-convex problem. Teleporting increases the gradient norm by two orders of magnitude over the standard initialization.

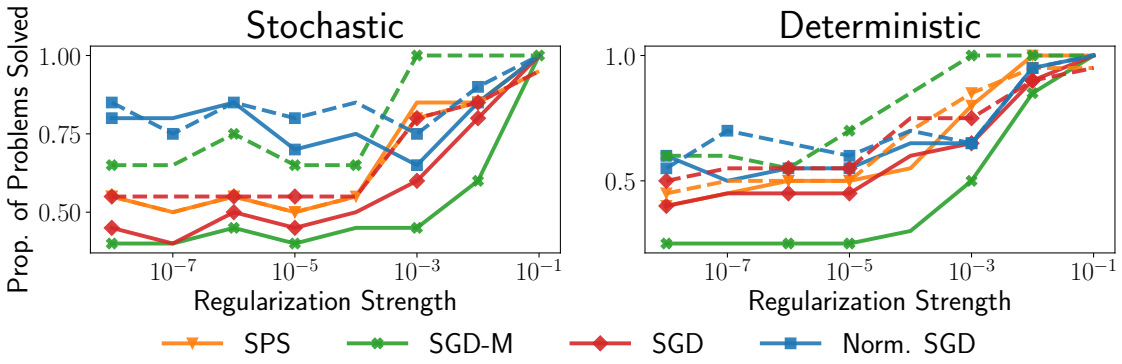

Figure 6: Effect of regularization strength on optimization when training three-layer ReLU networks on 20 datasets from the UCI repository. A problem is solved when $(f(w_k) - f(w^*))/f(w^*) \leq 0.5$, where $f(w^*)$ is the smallest objective found by any method. Methods initialized with teleportation (solid lines) are more sensitive to small regularization than methods without (dashed).

that decreasing the strength of weight-decay regularization negatively affects the success rate of teleportation compared to standard initializations. Thus, teleportation is best suited to problems where large regularization is already desirable for modeling reasons.

Stochastic Performance Profile: Fig. 7 provides a version of the performance profile from the main paper (Fig. 3) where all optimization methods are run with mini-batching. We observe similar results to the deterministic case, although the effect of teleportation does appear to be minimized when using stochastic methods.

MNIST and Fashion MNIST: Fig. 8 replicates our experiments on MNIST with stochastic gradients, while Fig. 9 shows that similar results hold for the Fashion MNIST dataset. Note that in both these experiments we try increasing the step-size 10 iterations after teleporting (deterministic case only) to test the hypothesis that overly small step-sizes are the case of slow convergence. We find that increasing the step-size leads to noisy updates and does not improve optimization speed.

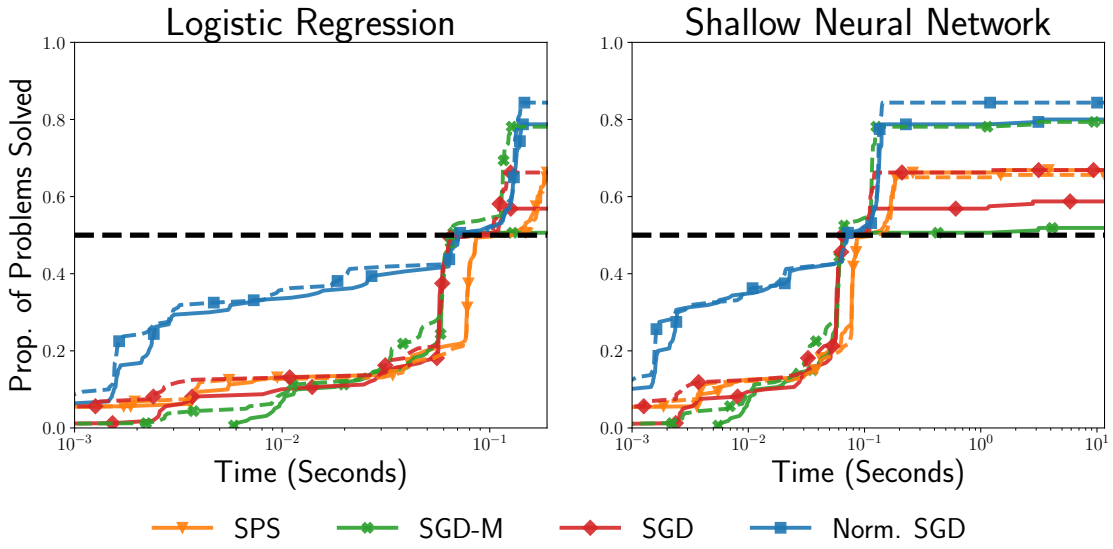

Figure 7: Performance profile comparing optimization methods with (solid lines) and without (dashed lines) initialization by sub-level set teleportation. A problem is solved when $(f(w_k) - f(w^*))/f(w^*) \leq 0.5$, where $f(w^*)$ is the smallest objective found by any method. All methods are run with mini-batches of size 64. The performance of methods with and without teleportation is generally similar, although the standard initialization performs slightly better overall.

Utility of Teleportation: Finally, Fig. 10 shows three UCI datasets where teleportation helps significantly over the standard initialization. We include these results to provide a balanced view on teleportation, the utility of which depends on the particular dataset and model considered.

C.2. Experimental Details

In this section we include additional details necessary to replicate our experiments. We run our experiments using PyTorch (Paszke et al., 2019). Unless otherwise stated, all experiments using neural networks are conducted on ReLU networks with two hidden layers, each of which has 100 units. When selecting step-sizes for optimization methods, we perform a grid-search using the grid $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. All experiment results are shown for three random restarts excepting the performance profiles, where averaging is not straightforward. We plot the median and first/third quartiles. Step-sizes are selected by minimizing the training loss at the end of the last epoch. For our teleportation method, we initialize the step-size at $\rho = 0.1$ and use the tolerances $\epsilon = \delta = 10^{-10}$. In practice, we scale γ_t by 2 for stability and relax the Armijo progress criterion with parameter $\alpha = 10^{-3}$. For SGD with momentum, we use the momentum parameter $\beta = 0.9$ and dampening parameter $\mu = 0.9$. We estimate f^* with zero for SPS.

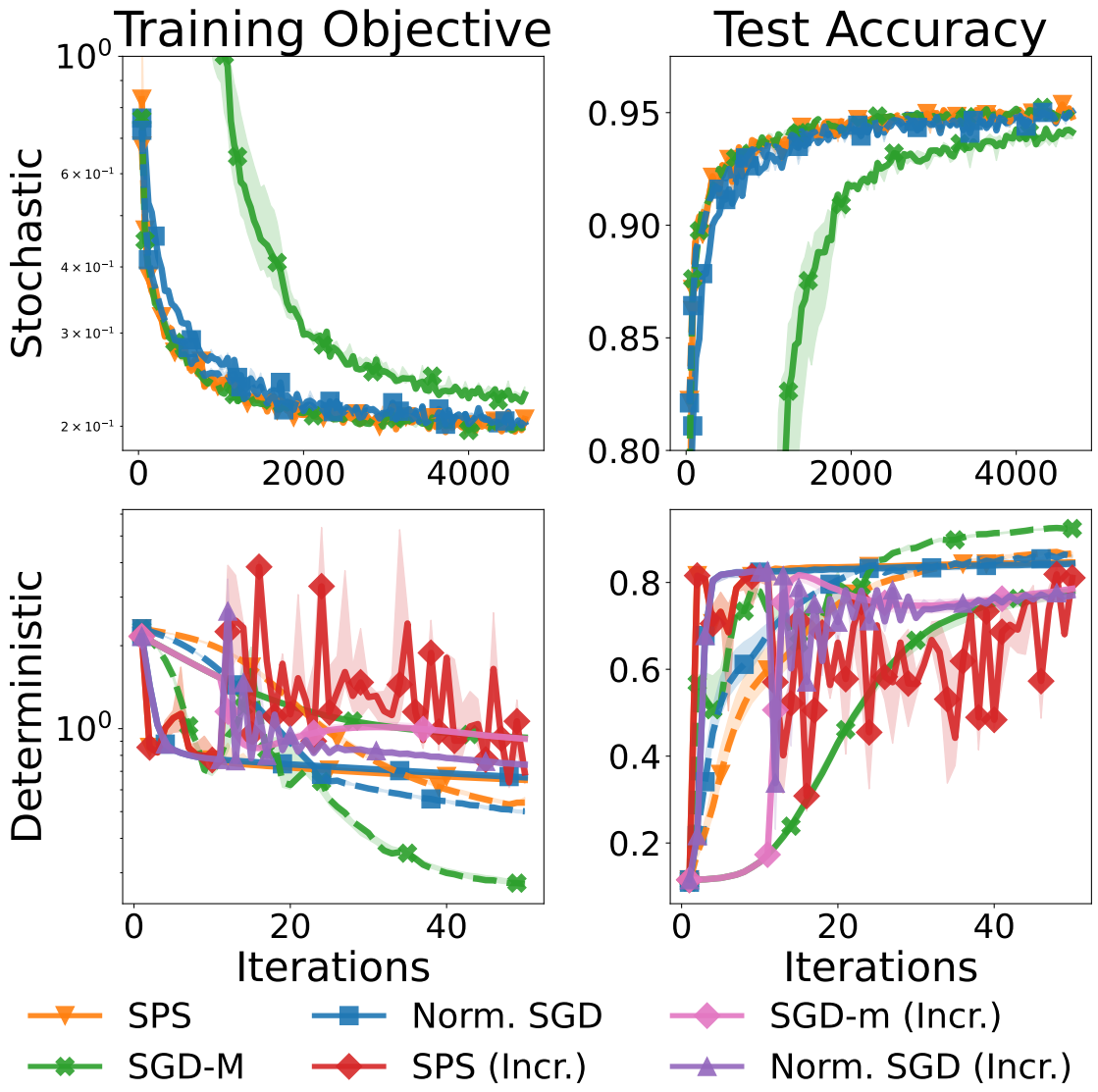

Figure 8: Performance of optimizers with (solid) and without (dashed) initialization by teleportation on MNIST. We train a ReLU MLP with two hidden layers of size 100. Stochastic methods are run with a batch-size of 128. In the deterministic setting, we try increasing the step-size of methods with teleportation by a factor of 10 after 10 epochs to see if slow convergence is due to overly small step-sizes. Larger step-sizes do not improve optimization speed, which indicates the "stalling" behavior we observe may be due to local geometry.

Test Functions: We use open-source implementations of the Booth and Goldstein-Price functions[1]. Gradient descent with and without teleportation are run with an Armijo line-search starting from step-size $\eta = 1$. Newton's method is run with a fixed step-size $\eta = 0.8$.

1. Available here: https://github.com/AxelThevenot/Python_Benchmark_Test_Optimization_Function_Single_Objective

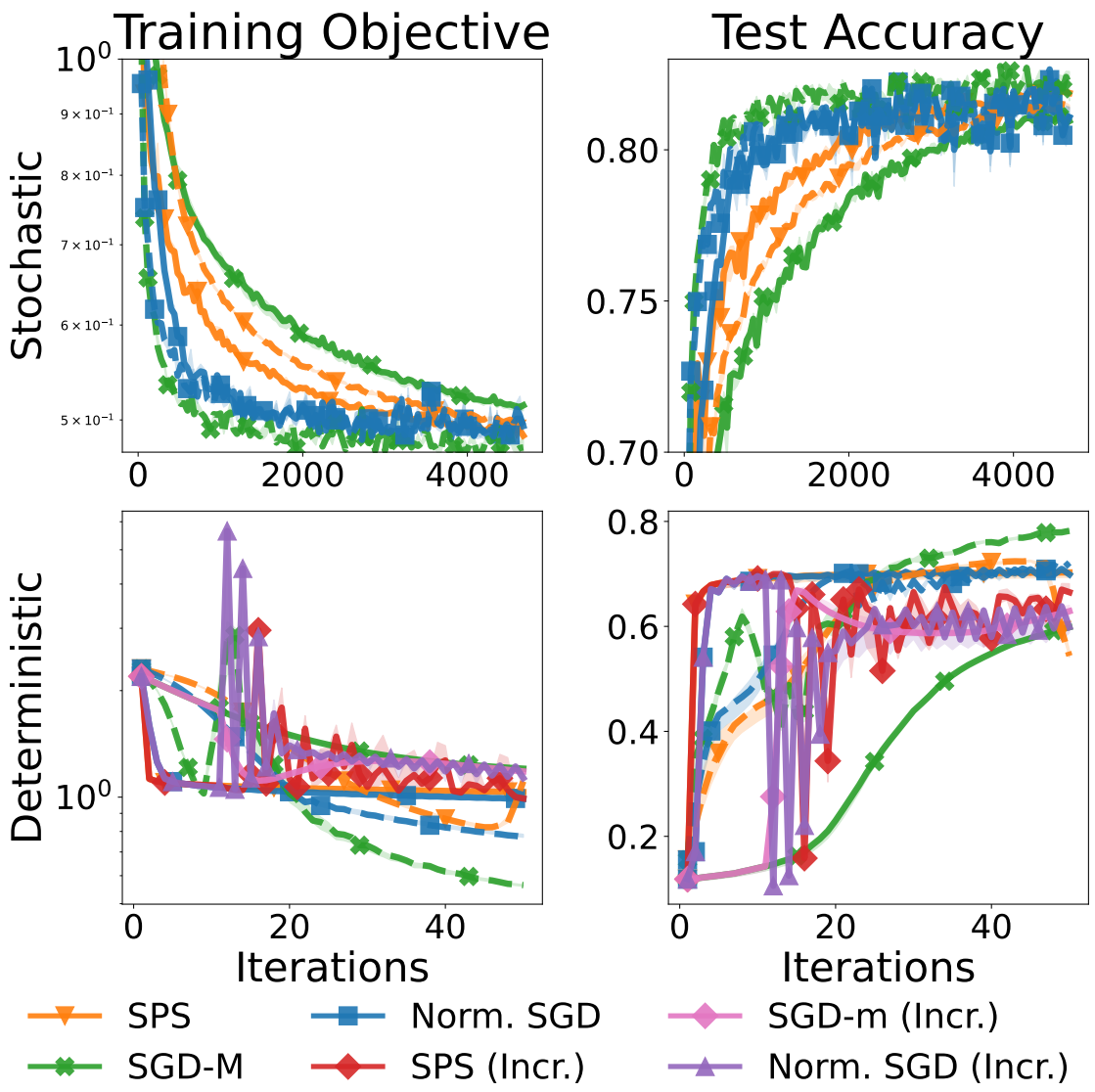

Figure 9: Performance of optimizers with (solid) and without (dashed) initialization by teleportation on Fashion MNIST. We follow the same experimental protocol as in Fig. 8 and observe similar results.

Teleportation on MNIST: We use a two-layer ReLU network with fifty hidden units. The strength of weight decay regularization is set at $\lambda = 1.8$.

UCI Performance Profiles: We run on the following 20 binary classification datasets selected from the UCI repository: `blood`, `breast-cancer`, `chess-krvkp`, `congressional-voting`, `conn-bench-sonar`, `credit-approval`, `cylinder-bands`, `hill-valley`, `horse-colic`, `ilpd-indian-liver`, `ionosphere`, `magic`, `mammographic`, `musk-1`, `ozone`, `pima`, `tic-tac-toe`, `titanic`, `ringnorm`, `spambase`. We use the pre-processed datasets provided by Fernández-Delgado et al. (2014), although we do not use their splits since these are known to

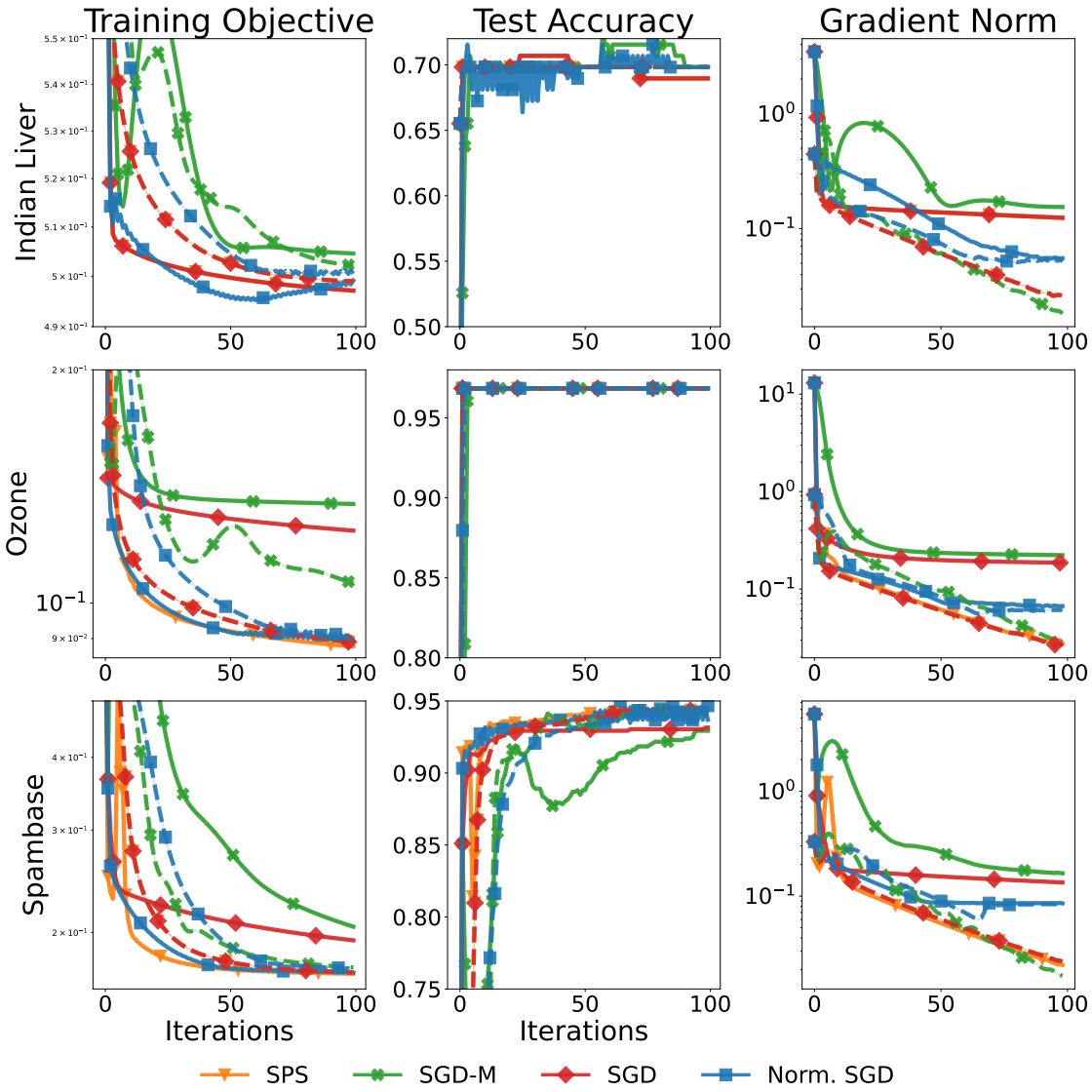

Figure 10: Performance of methods initialized using sub-level set teleportation for training three layer ReLU networks on three datasets from the UCI repository. These results are specially selected from the 20 datasets used in the performance profile (Fig. 3) to showcase situations where teleportation outperforms the standard initialization. All methods are run in batch mode. Normalized SGD and SPS perform particularly well with teleportation and do not show the stalling behavior observed on MNIST.

have test set contamination. To obtain 180 distinct problems, we also consider regularization parameters from the grid $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$. For the stochastic setting, we use batch-sizes of 64.

Image Classification: We use a fixed strength of $\lambda = 10^{-2}$ for the weight decay regularization. All other settings are as described above.

Effects of Regularization: The data for Fig. 6 comes directly from the performance profiles in Fig. 3 and Fig. 7. All results are shown for ReLU networks.

Additional UCI Plots: The data for Fig. 10 comes directly from the performance profile in Fig. 3. All results are shown for regularization parameter $\lambda = 0.2$.