Surfing: Iterative Optimization Over Incrementally Trained Deep Networks

Ganlin Song, Zhou Fan, John Lafferty
Yale University

Motivation

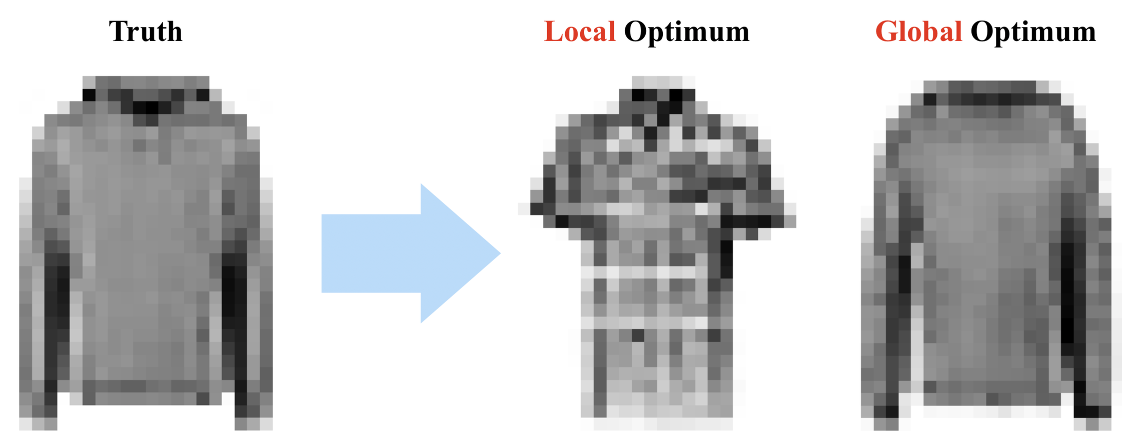
We minimize

$$f_\theta(x) = \frac{1}{2}\|AG_\theta(x) - Ay\|^2$$

$A \in \mathbb{R}^{m\times n}$ matrix, $y \in \mathbb{R}^n$ the true signal ($m \ll n$), G_θ generative network.

Goal: Find the minimizer $\hat{x}$ and recover y by $G_\theta(\hat{x})$.

PROBLEM: $f_\theta(x)$ is usually non-convex; gradient descent cannot find global minimum.

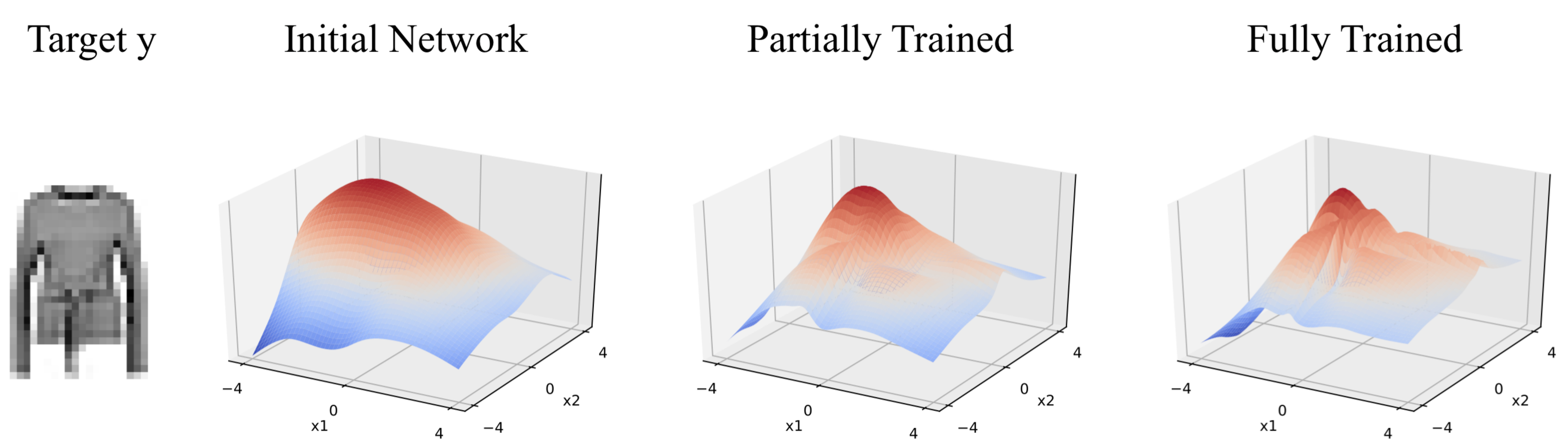


OUR RESULTS:

- Propose a novel algorithm that outperforms gradient descent
- Theoretical analysis that ensures convergence under certain conditions

Surfing Algorithm

Landscape of $f_\theta(x)$ is "nice" at initilization and slowly becomes nonconvex and "wavy" as weights θ of G_θ are trained



Main idea: Use intermediate networks from training process and track the global optimum

Algorithm:

Given a sequence of networks $G_0, G_1, \dots, G_T$, write

$$f_t(x) = \frac{1}{2}\|AG_t(x) - Ay\|^2, \quad \forall\, t \in [T].$$

Optimize f_0 to obtain the minimizer x_0, then apply gradient descent on f_t for $t = 1, 2, \dots, T$ iteratively, initializing at the minimizer x_{t-1}.

Theoretical Results

Consider G with ReLU activation,

$$G(x,\theta) = V\sigma(W_d \dots \sigma(W_2\sigma(W_1 x + b_1) + b_2)\dots + b_d).$$

1. If θ_0 is Gaussian and G is sufficiently expansive, then all critical points of $f_0(x)$ belong to a small neighborhood around 0 (with high probability). Builds on Hand and Voroninski (2017).

   Proof technique:
   - G is piecewise linear, gradient of f_0 has explicit expression
   - Concentration bounds give $\nabla f_0(x) = 2^{-d}x + O(\varepsilon(1+\|x\|))$
   - Find descent direction v such that $D_v f_0(x) < 0$ for all $\|x\| > O(\varepsilon)$

2. Consider a network flow $G^s(x) = G(x, \theta(s))$ for $s \in [0, S]$ and corresponding objective $f^s(x) = f(x, \theta(s))$. If the weights of G^s are bounded and the global minimizer of f^s is unique and Lipschitz-continuous, then the projected-gradient surfing can keep track of the minimizer with a small time discretization step δ of G^s.

   Projected-gradient surfing:
   - Identify all the linear pieces $\{P_1, \dots, P_l\}$ for current G_t that could contain global minimizer of f_t
   - Apply projected gradient descent $x \leftarrow \mathrm{Proj}_P(x - \eta\nabla f_t(x))$ for each P_i

Experiments

1. **Inverting the generator**: $f(x) = \frac{1}{2}\|G(x) - G(x_*)\|^2$

   Compared with regular gradient descent (ADAM), surfing has
   - Higher success rate
   - Comparable computational cost

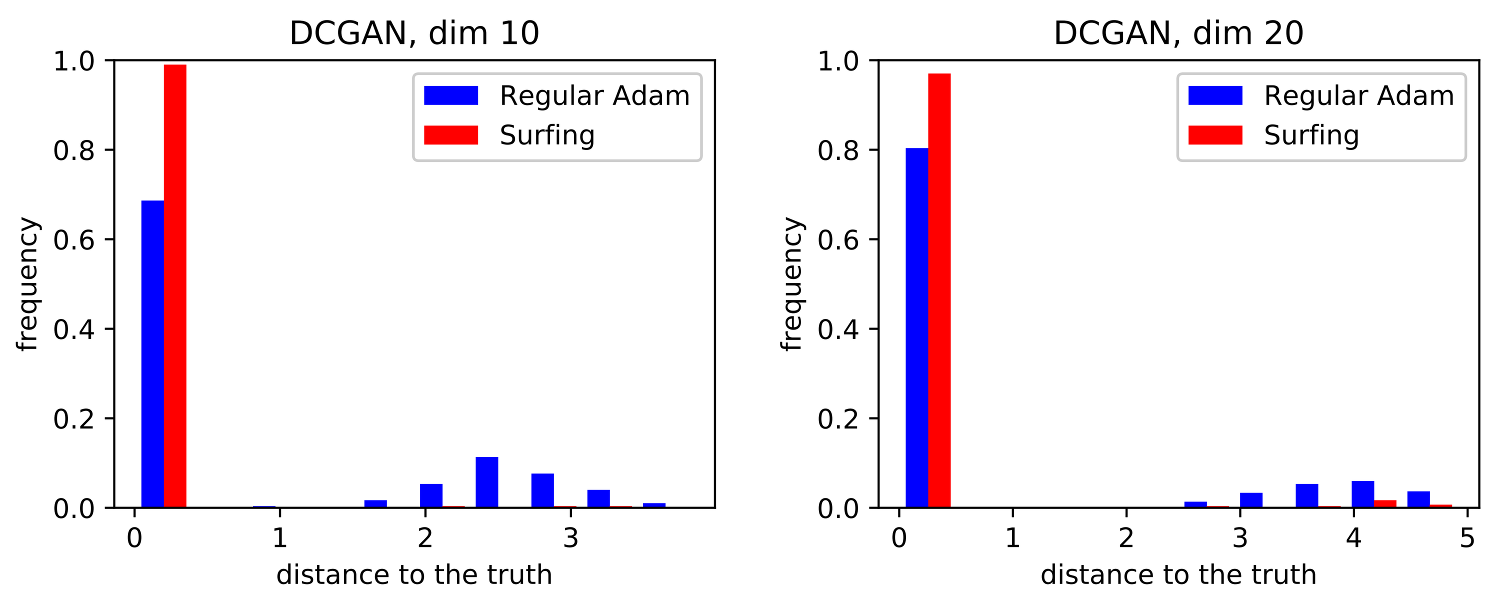


Figure 1: Distribution of distance between solution $\hat{x}$ and the truth x_*

Input dimension		5	10	20	5	10	20	5	10	20
	Model	DCGAN			WGAN			WGAN-GP		
% successful	Regular Adam	48.3	68.7	80.0	56.0	84.3	90.3	47.0	64.7	64.7
	Surfing	78.3	98.7	96.3	81.7	97.3	99.3	83.7	95.7	97.3
# iterations	Regular Adam	618	4560	18937	464	1227	3702	463	1915	15445
	Surfing	741	6514	33294	547	1450	4986	564	2394	25991

Table 1: Percentages of solutions $\hat{x}$ satisfying $\|\hat{x} - x_*\| < 0.01$

2. **Compressed sensing, y in the range of G**: $f(x) = \frac{1}{2}\|AG(x) - AG(x_*)\|^2$

   See Bora, Jalal, Price, & Dimakis (2017)

   Surfing finds global optimum more consistently than direct GD

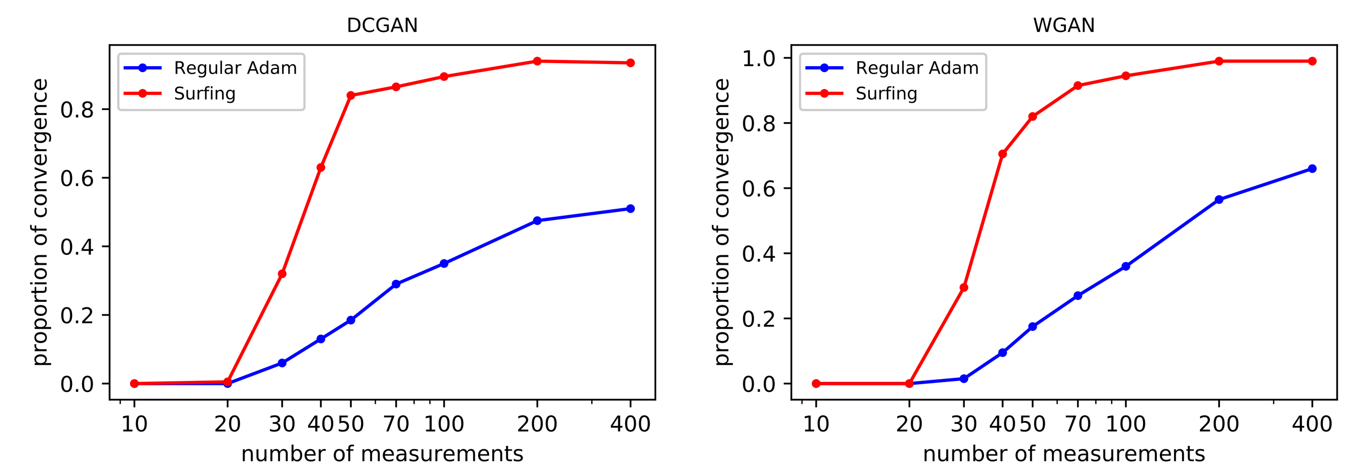


Figure 2: Compressed sensing setting for exact recovery

3. **Compressed sensing, y from test data**: $f(x) = \frac{1}{2}\|AG(x) - Ay\|^2$

   Surfing gives favorable reconstruction of signal y

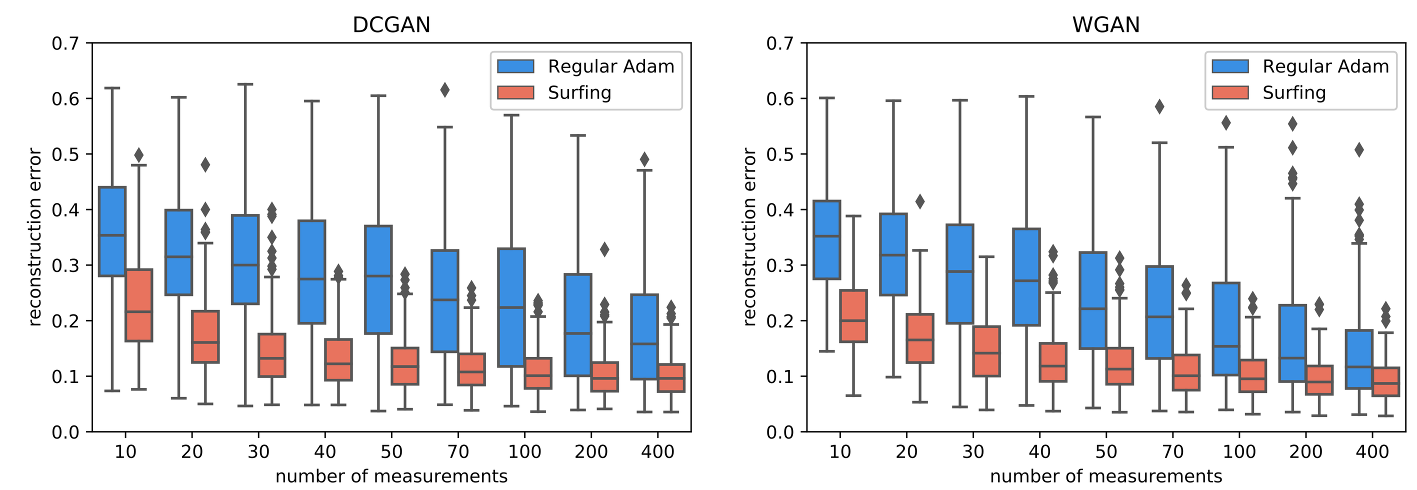


Figure 3: Compressed sensing setting for approximation (rate-distortion)

The reconstruction error is $\sqrt{\frac{1}{n}\|G(\hat{x}) - y\|^2}$.

Future Work

1. Bridging gap between practice and theory—we use projected-gradient surfing only in our analysis.
2. Objective function with trained network often does have a nice landscape. Why?
3. Can we constrain or regularize the training process of G so that simple gradient descent is guaranteed to succeed?
4. Can the idea behind surfing be applied to other optimization problems in machine learning?