1 Background

1.1 Towards the Time- and Energy-Efficient, Optimal Neural Network Design

1.2 State-of-the-Art Fine-tuning Methods in Quantization-Aware Training Procedures

1.2.1 BinaryConnect

1.2.2 Straight-Through Estimators

1.3 Formalization of the Optimization Problem and Inherent Difficulty

2 Preliminaries

In this section, we will consider the following optimization problem:

$$\min_{x \in C} \ell(x), \text{where } C := \{x \in \mathbb{R}^n \mid g(x) \geq 0\} \tag{P}$$

with the objective function $f : \mathbb{R}^n \to \mathbb{R}$, and constraints described by the function $g : \mathbb{R}^n \to \mathbb{R}^d$.

2.1 Muehlebach-Jordan's Work on Local Approximation of the Constraint Set

Assumption 1. The functions ℓ, g are continuously differentiable and have an L-Lipschitz continuous gradient. Additionally, ℓ is lower-bounded.

Assumption 2. The Mangasarian-Fromovitz constraint qualification (MFCQ) is satisfied for all $x \in \mathbb{R}^n$, i.e. $\forall x \in \mathbb{R}^d$, $\exists w \in \mathbb{R}^n$ s.t. $\nabla g_i(x) w > 0$ for all $i \in I(x)$, where $I(x) = \{i \in \mathbb{Z} \mid g_i(x) \leq 0\}$.

Algorithm 1 Muehlebach-Jordan's First-Order Gradient Flow Algorithm

1: Start with an initial dual guess $x_0 \in \mathbb{R}^d$
2: **for** $k = 0, 1, \ldots$ **do**
3: Let $-R_k \in \partial \psi_{V_\alpha(x_k)}((x_{k+1} - x_k)/T)$
4: $x^{(k+1)} \leftarrow x^{(k)} - T \nabla f(x_k) + T R_k$
5: **end for**

Additionally, Muehlebach and Jordan (2022) have shown the equivalence between $-\nabla f(x_k) + R_k$ and

$$\operatorname*{argmin}_{v \in V_\alpha(x)} (1/2) \|v + \nabla f(x)\|^2 \tag{1}$$

due to Assumption 2. We emphasize this formulation for its characterization of the behavior of the above algorithm (and the algorithms developed in this paper).

2.2 Constructing a Constrained Quantization Function

Leconte et al.'s (2023) approach to the optimization problem relaxes the quantization constraints to a set of "smoothed" interval constraints. They first create a continuous function that quantifies "how far away" a parameter is from its quantization levels.

Definition 1. Let $\varepsilon \in [0, 1]$, $w^i \in \mathcal{Q}^i, i \in \{1, 2, \ldots, d\}$, i.e. each w^i coordinate-wisely takes value

from its own set of quantization levels $\mathcal{Q}^i = \{q_1^i, q_2^i, \ldots, q_{K^i}^i\}$. We define the piecewise function

$$\psi_\varepsilon^i(w^i) := \begin{cases} \varepsilon - (q_1^i - w^i)^2, & w^i < q_1^i, \\ \varepsilon - (w^i - q_{j-1}^i)^2(w^i - q_j^i)^2, & q_{j-1}^i \le w^i < q_j^i, j = 2, \ldots, K, \\ \varepsilon - (w^i - q_K^i)^2, & w^i \ge q_{K^i}^i, \end{cases}$$

for all $w^i \in \mathbb{R}$.

Assumption 3. The stepsizes $(\gamma_k)_{k\ge 0}$ are positive, $\sum_{k=0}^{\infty} \gamma_k = \infty$ and $\lim_{k\to\infty} \gamma_k = 0$.
Assumption 4. $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$.

We can simplify the tangent cone and the normal cone for any $x \in C$ as follows, due to the Mangasarian-Fromovitz constraint qualification of g:

$$T_C(x) = \{x \mid \nabla g_i(x)^\top x \ge 0, \forall i \in I_x\}, N_C(x) = \left\{ \lambda \in \mathbb{R}_+^d \middle| - \sum_{i\in I(x)} \lambda_i \nabla g_i(x) \right\}.$$

Definition 2. Further define the set $V_{\varepsilon,\alpha}(x) := \{v \in \mathbb{R}^n \mid \nabla g_i(x)^\top v + \alpha g_i(x) \ge 0, \forall i \in I_x\}$, where $\alpha > 0$.

Condition A. $V_\alpha(w)$ is empty if and only if there is $1 \le i \le d$ such that $w^i = (q^i_{Q^i(w^i)} + q^i_{Q^i(w^i)+1})/2$.

Indeed. For the case where $x \in C$, $V_\alpha(x)$ is nothing but $T_C(x)$. Otherwise, consider any $g_i(x) < 0$ for some i, by MFCQ there exists u such that $\forall i \in I_\varepsilon(x), \nabla g_i(x)^\top u > 0$ and therefore by scaling we have $\nabla g_i(x)^\top v \ge -\alpha g_i(x) \ge 0$.

We consider following formulation for the update direction again, and observe that under the current settings,

$$v_k = \operatorname*{argmin}_{v\in V_{\varepsilon,\alpha}(w_k)} (1/2)\|v + \widehat{\nabla}\ell(w_k)\|^2$$

admits an explicit solution

$$[s_{\varepsilon,\alpha}(\widehat{\nabla}\ell(w_k), w_k)]^i = \begin{cases} -\widehat{\nabla}\ell(w_k^i), & \text{if } \psi_\varepsilon(w^i) > 0 \text{ or } -\psi_\varepsilon'(w^i)\widehat{\nabla}\ell(w_k^i) \ge -\alpha\psi_\varepsilon(w^i) > 0, \\ \mathrm{clip}(-\alpha\psi_\varepsilon(w^i)/\psi_\varepsilon'(w^i), M_\varepsilon), & \text{otherwise,} \end{cases}$$

for w such that $w^i \ne (q^i_{Q^i(w^i)} + q^i_{Q^i(w^i)+1})/2$, where $Q^i(w^i)$ is the unique index satisfying $q^i_{Q^i(w^i)} \le w^i < q^i_{Q^i(w^i)+1}$.

Algorithm 2 The `ASkewSGD` Algorithm

1: Select a sequence (γ_k) of step sizes, and size of the mini-batch $N_b \le N$.
2: Start with an initial dual guess $x_0 \in \mathbb{R}^d$
3: **for** $k = 0, 1, \ldots$ **do**
4: Sample a mini-batch of N_b observations $\{j_1, j_2, \ldots, j_{N_b}\}$ in $\{1, 2, \ldots, N\}$
5: Compute the stochastic gradient $\widehat{\nabla}\ell(w_k) = 1/N_b \sum_{i=1}^{N_b} \nabla\ell(w_k; \xi_{j_i})$
6: Compute the update direction $v_k = s_{\varepsilon,\alpha}(\widehat{\nabla}\ell(w_k), w_k)$
7: $x^{(k+1)} \leftarrow x^{(k)} - \gamma_k v_k$
8: **end for**

2.3 Unconstrained Minimax Optimization by Lagrangian Multipliers

3 Convergence Proof

In this section, we follow the work of Muehlebach and Jordan's (2022) to analyze the convergence of the deterministic version of `ASkewSGD` under our new construction for the constraint set C_ε.

The novelty of our new setup has successfully extended the constructions by Leconte et al. (2023) toward better properties including piecewise convexity, and constraints separation for every quantization level etc. This fact is helpful if we need to characterize the smoothness[1] of the Lagrangian dual of (P)

Definition 3. We define the piecewise function

$$\hat{\phi}^i(w^i) := \begin{cases} (q_1^i - w^i), & w^i < q_1^i, \\ (q_{j-1}^i - w^i)(w^i - q_j^i), & q_{j-1}^i \le w^i < q_j^i, j = 2, \dots, K, \\ w^i - q_K^i, & w^i \ge q_{K^i}^i, \end{cases}$$

for all $w^i \in \mathbb{R}$.

Observing the convexity of ψ, it would be natural to think about whether we treat every convex piece independently.

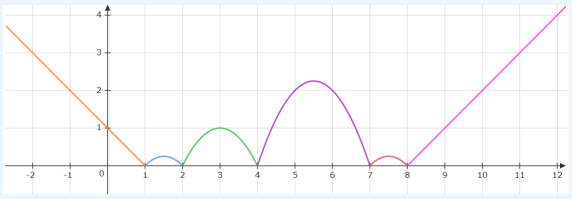
Definition 4. A function $F : \mathbb{R}^n \mapsto \mathbb{R}$ is called a piecewise convex function if it can be decomposed into:

$$F(x) = \min\{f_1(x), f_2(x), \dots, f_m(x)\},$$

where $f_j : \mathbb{R}^n \mapsto \mathbb{R}$ are convex functions for all $j \in [M]$.

Lemma 1. Let $\varepsilon \in [0, 1]$, $\hat{\psi}^i(w^i) := \varepsilon - \hat{\phi}^i(w^i)$ is a piecewise convex function.

Proof. That is, to prove that $\hat{\psi}$ can be decomposed into concave functions $f_1, \dots, f_m$ where $F(x) = \max(f_1(x), f_2(x), \dots, f_m(x))$.



Select $m = K + 1$ such that the f_i's $(i = 1, \dots, K + 1)$ corresponds to the analytic continuation of every piece of the piecewise function:

$$\begin{cases} f_1^i(w^i) := (q_1^i - w^i), \\ f_j^i(w^i) := (q_{j-1}^i - w^i)(w^i - q_j^i), j = 2, \dots, K^i, \\ f_{K^i+1}^i(w^i) := w^i - q_{K^i}^i. \end{cases}$$

The proof is pictorially shown by the plot above. Observe that any $f_j^i > 0$ implies $f_k^i < 0$ for any $j \neq k$, this is guaranteed by the choice of f_is (being linear or quadratic functions). □

Corollary 1. The constraint on the i-th parameter w^i described by $g_i := \{w^i : \hat{\psi}^i(w^i) \ge 0\}$ can be replaced by $g_i' = (f_1^i, \dots, f_{K^1}^i) \succeq 0$.

[1] Note that in the context of optimization, smoothness controls the changes in gradients.

This corollary is direct from the definition of piecewise convex function. For the purpose of simplicity, we first adopt the original constraint functions described by ϕ^i instead of f_j^i. We will explicitly state when we apply the replacement.

Theorem 1. Under Assumption 1 and 3 and $0 < \varepsilon \leq \inf_{1\leq i\leq d} \inf_{1\leq j\leq K^i} \left|c_j^i - c_{j+1}^i\right|^2/4, \alpha \leq L$, where $\{c_j^i\}$ are the quantization levels. Then, $\ell(w_k)$ converges and $\lim_{k\to\infty} d(w_k, \mathcal{Z}_\varepsilon) = 0$ almost surely.

Again, we have followed the journey of Muehlebach and Jordan's (2022) work. Their convergence proof hinges heavily on the properties that:

- Once w_k has fallen into a convex region C, the multiplier λ_k is feasible for the dual (1) at time $k+1$.
- The function $\mathcal{L}(x_k, \lambda_k) - \|\nabla_x \mathcal{L}(x_k, \lambda_k)\|^2/2\alpha$ increases from x_k to x_{k+1} (for a fixed λ_k) due to the smoothness of ℓ.

The proof of Theorem 1 is done by several smaller lemmata. First, we bounding the step size such that the iterates fall into a convex region. Then, we show that the corresponding dual of (1) $d(x_k)$ is monotonically increasing and bounded above by f^*. Finally, we expand $d(x_{k+1})$ as a telescoping sum, and it immediately follows that v_k converges to zero for large k.

Lemma 2. Strong duality holds for (1) for every $x \in \mathbb{R}^n$ when $V_{\alpha,\varepsilon}$ is non-empty (See Condition A). For $\alpha \geq 0$, the dual can be rewritten as

$$\max_{\lambda\succeq 0} -\frac{1}{2}\|\lambda^\top \nabla g(x) - f(x)\|^2 - \alpha\lambda^\top g(x) \tag{2}$$

Proof. Let $D_x : \{\lambda \in \mathbb{R}^n : \lambda_i = 0 \text{ if } i \notin I(x)\}$. For clarity, we reformulate the primal problem in Lagrangian:

$$\min_{v\in\mathbb{R}^n} \max_{\lambda\in D_x} \frac{1}{2}\|v + \nabla f(x)\|^2 - \lambda^\top(\nabla g(x)^\top v + \alpha g(x)). \tag{3}$$

Then, the corresponding dual is as follows:

$$\max_{\lambda\in D_x} \min_{v\in\mathbb{R}^n} \frac{1}{2}\|v + \nabla f(x)\|^2 - \lambda^\top(\nabla g(x)^\top v + \alpha g(x)), \tag{4}$$

which then resolves into (2) by picking $v^* = \lambda^\top \nabla g(x) - \nabla f(x)$ and the addition of the constant term $-(1/2)\|\nabla f(x)\|^2$.

We show that Slater's condition holds for $V_{\varepsilon,\alpha}$, i.e. there exists a $v \in \mathbb{R}^n$ such that $\nabla g_i(x)^\top v + \alpha g_i(x) > 0$ for all $i \in I_\varepsilon(x)$.

Let $\bar{v} \in \mathbb{R}^n$ satisfy $\nabla g_i(x)\bar{v} + \alpha g_i(x) = 0, \forall i \in I_\varepsilon(x)$. By MFCQ, there exists a $w \in \mathbb{R}^n$ such that $\nabla g_i(x)w > 0, \forall i \in I_\varepsilon(X)$. Picking $v = \bar{v} + \xi w$ for some $\xi > 0$ satisfies the required condition.

By the fact that (1) is convex, □

Fact. The MFCQ condition is automatically satisfied for $x \in C_\varepsilon$ by construction.

Tangent cones have a natural role in the theory of flow-invariant sets and gradient inclusions.

Definition 5. The Clarke's tangent cone of C contains all $\delta x \in T_C(x)$ if there exists two sequences $x_j \to x, x_j \in C, t_j \downarrow 0$ such that $(x_j - x)/t_j \to \delta x$. The normal cone is defined as follows: $N_C(x) = \{\lambda \in \mathbb{R}^n \mid \lambda^\top \delta x \leq 0, \forall \delta x \in T_C(x)\}$.

Lemma 3. Suppose that $x \in C$, then every $\delta x \in T_C(x)$ satisfies $\nabla g_i(x)\delta x \geq 0, \forall i \in I_x$. The converse also holds.

Proof. $(\Rightarrow)$: $\delta x \in T_C(x)$ implies that there exists two sequences $\{x_j\} \to x, \{x_j\} \subset C, t_j \downarrow 0$ for all $j \in \mathbb{N}$ and

$$\frac{x_j - x}{t_j} \to \delta x,$$

which implies that

$$\frac{g(x_j) - g(x)}{x_j - x} \cdot \frac{x_j - x}{t_j} \geq 0.$$

This is because $x_j \in C$ implies that $g_i(x_j) \geq 0$ and $g_i(x) \leq 0$ for all $i \in I(x)$.

$(\Leftarrow)$: Adapted from R. Herzog, 2023, a simplified version. Let δx satisfy $\nabla g_i(x)\delta x \geq 0, \forall i \in I_x$, also let δy be given by MFCQ such that $\nabla g_i(w)\delta y > 0, \forall i \in I(x)$. Put $\ell(t) := \delta x + t \cdot \delta y$. Then for all $t > 0$, we have $\nabla g_i(x)\ell(t) > 0, \forall i \in I(x)$, implying that $\ell(t)$ are all feasible MFCQ vectors.

Now, we claim that $\ell(t) \in T_C(x)$ for all $t \in \mathbb{R}_{++}$. Let $\gamma(t) := x + t\ell(t)$, $t \in (-\varepsilon, \varepsilon)$, for an infinitestimally small ε, given by the continuity of g. Then, $y(t) \in C$ for every $t \in [0, \varepsilon)$ and $\gamma(0) = x, \gamma'(0) = \ell(t)$. For an arbitrary sequence $\{t_j\} \downarrow 0$ and $x_k = \gamma(t_j) \to x$ we have

$$\ell(t) = \gamma'(0) = \lim_{j\to\infty} \frac{\gamma(t_j) - \gamma(0)}{t_j - 0} = \lim_{j\to\infty} \frac{x_j - x}{t_j} \in T_C(x).$$

Since $T_C(x)$ is closed, $\delta x = \lim_{t\to 0} \ell(t) \in T(x)$. □

Now, we can simplify the tangent cone and the normal cone for any $x \in C$ as follows, due to the Mangasarian-Fromovitz constraint qualification:

$$T_C(x) = \{x \mid \nabla g_i(x)^\top x \geq 0, \forall i \in I_x\}, N_C(x) = \left\{ \lambda \in \mathbb{R}^d_+ \middle| - \sum_{i\in I(x)} \lambda_i \nabla g_i(x) \right\}.$$

Definition 6. Further define the set $V_\alpha(x) := \{v \in \mathbb{R}^n \mid \nabla g_i(x)^\top v + \alpha g_i(x) \geq 0, \forall i \in I_x\}$, where $\alpha > 0$. $V_\alpha(x)$ is guaranteed to be non-empty for any x.

Indeed. For the case where $x \in C$, $V_\alpha(x)$ is nothing but $T_C(x)$. Otherwise, consider any $g_i(x) < 0$, by MFCQ there exists u such that $\nabla g_i(x)^\top u > 0$ and therefore by scaling we have $\nabla g_i(x)^\top v \geq -\alpha g_i(x) \geq 0$.

Definition 7. The indicator function for a set C is defined as:

$$\psi_C(x) = \begin{cases} 0, & x \in C, \\ \infty, & \text{otherwise.} \end{cases}$$

Theorem 2. Let $x : [0, \infty) \to \mathbb{R}^n$ be an absolutely continuous trajectory with a piecewise continuous derivative. Then, for any $x(0) \in C$, the following are equivalent:

$$\begin{aligned}
&\dot{x}(t) := -\nabla f(x(t)) + R(t), -R(t) \in N_C(x(t)), && \forall t \in [0, \infty) \text{ almost everywhere,} \\
&\dot{x}(t)^+ := -\nabla f(x(t)) + R(t), -R(t) \in \partial \psi_{V_\alpha(x(t))}(\dot{x}(t)^+), && \forall t \in [0, \infty), \\
&\dot{x}(t)^+ := -\operatorname*{argmin}_{v \in V_\alpha(x(t))} \frac{1}{2} |v + \nabla f(x(t))|^2, && \forall t \in [0, \infty).
\end{aligned}$$

Lemma 4. Using the `ASkewSGD` algorithm with step sizes $\{\gamma_k\}$ of $\sum_{i=1}^{\infty} \gamma_i = \infty$, $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$, the iterate $\{w_k\}$ is guaranteed to converge and $\lim_{k \to \inf} d(w_k, C_\varepsilon) = 0$.

Proof. See Leconte et al., 2023, Appendix A.3. □

Lemma 5. Let $k_0 = \sup_{1 \le i \le d, 1 \le j \le K_i} \sup\{k : \gamma_k M \ge \max(c_- - \frac{c_j^i + c_{j+1}^i}{2}, -c_+ + \frac{c_j^i + c_{j+1}^i}{2})\}$. Since w must

4 Lagrange Duality

The problem $\mathcal{P}$ of

$$\min_{x \in C} f(x), C = \{x \in \mathbb{R}^n \mid g(x) \leq 0\}$$

is equivalent to the primal problem

$$\inf_{x \in \mathbb{R}^n} \sup_{\lambda \succeq 0} f(x) + \sum_{i=1}^{d} \lambda_i g_i(x).$$

We consider the dual problem

$$\sup_{\lambda \succeq 0} \inf_{x \in \mathbb{R}^n} f(x) + \sum_{i=1}^{d} \lambda_i g_i(x).$$

Theorem 3. Suppose that x^* is a local minimizer of $\mathcal{P}$ which satisfies the MFCQ. Then there exist Lagrange multipliers λ^* (not necessary unique) such that the KKT conditions are satisfied. The set of Lagrange multipliers $\Lambda(x^*)$ is compact.

Therefore, the KKT points can be captured by the following set:

$$\mathcal{Z}_\varepsilon = \{w \in C_\varepsilon : 0 \in -\nabla \ell(w) + N_{C_\varepsilon}(w)\}$$

Theorem 4. If $f : \mathbb{R}^d \mapsto \mathbb{R}$ is twice continuously differentiable and satisfies the strict saddle property, then gradient descent with a random initialization and sufficiently small constant step size converges to a local minimizer or negative infinity almost surely. Call x a critical point of f if $\nabla f(x) = 0$, and say that f satisfies the strict saddle property if each critical point x of f is either a local minimizer, or a "strict saddle", i.e, $\nabla^2 f(x)$ has at least one strictly negative eigenvalue.(J. D. Lee, in PMLT, 2016)

Algorithm 3 Dual gradient ascent method (convex constraints)

1: Start with an initial dual guess $\lambda(0) \geq 0$.
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad x^{(k)} \in \operatorname{argmin}_x \ell(x) + (\lambda^{(k-1)})^\top g(x)$
4: $\quad \lambda^{(k)} = \max\{\lambda^{(k-1)} + \gamma_k g(x^k), 0\}$
5: **end for**

Assumption 5. The objective function f is convex and continuously differentiable.

How we can find x^* efficiently, as $\nabla \ell$ is implicit?

The problem is that $f(x) - \lambda g$ is a combination of a convex and a concave function (where g is convex and $g \geq 0$ is required).

`https://proceedings.neurips.cc/paper_files/paper/2023/file/a961dea42c23c3c0d01b79918701fb6e-Paper.pdf`