Truncated Inference for Latent Variable Optimization Problems: Application to Robust Estimation and Learning Supplementary Material

Christopher Zach[0000−0003−2840−6187] and Huu Le[0000−0001−7562−7180]

Chalmers University of Technology, Gothenburg, Sweden
{zach,huul}@chalmers.se

1 Relaxed Generalized MM and the gradient method

We assume that $\overline{J}(\cdot;\overline{u})$ has a Lipschitz gradient with constant L, and therefore

$$\overline{J}(\theta;\overline{u}) \leq \overline{J}(\theta^0;\overline{u}) + \nabla\overline{J}(\theta^0;\overline{u})^T(\theta-\theta^0) + \frac{L}{2}\|\theta-\theta^0\|^2 \tag{1}$$

for all $\overline{u}$, θ and θ^0. We recall the relaxed GMM condition,

$$\begin{aligned}\overline{J}(\theta^{(t-1)};\overline{u}^{(t)}) &\leq \eta\underline{J}(\theta^{(t-1)},\underline{u}^{(t)}) + (1-\eta)\overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)}) \\ &= \overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)}) - \eta\underbrace{\left(\overline{J}(\theta^{(t-2)},\overline{u}^{(t-1)}) - \underline{J}(\theta^{(t-1)},\underline{u}^{(t)})\right)}_{=:c_t}.\end{aligned}$$

If $\theta^{(t-1)} = \theta^{(t-2)} - \frac{1}{L}\nabla\overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)})$, then

$$\begin{aligned}\overline{J}(\theta^{(t-1)};\overline{u}^{(t-1)}) &\leq \overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)}) + \nabla\overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)})^T(\theta^{(t-1)}-\theta^{(t-2)}) \\ &\quad + \frac{L}{2}\|\theta^{(t-1)}-\theta^{(t-1)}\|^2 \\ &= \overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)}) - \frac{1}{2L}\|\nabla\overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)})\|^2.\end{aligned} \tag{2}$$

Consequently,

$$\begin{aligned}c_t &= \overline{J}(\theta^{(t-2)},\overline{u}^{(t-1)}) - \underline{J}(\theta^{(t-1)},\underline{u}^{(t)}) \\ &\geq \overline{J}(\theta^{(t-1)};\overline{u}^{(t-1)}) + \frac{1}{2L}\|\nabla\overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)})\|^2 - \underline{J}(\theta^{(t-1)},\underline{u}^{(t)}) \\ &\geq J(\theta^{(t-1)}) - \underline{J}(\theta^{(t-1)},\underline{u}^{(t)}) + \frac{1}{2L}\|\nabla\overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)})\|^2 \\ &= \frac{1}{2L}\|\nabla\overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)})\|^2 \geq 0,\end{aligned} \tag{3}$$

where we first used Eq. 2 and then the lower and upper bounds on J. From Proposition 1 we know that $c_t \to 0$, which implies that $\|\nabla\overline{J}(\theta^{(t-2)};\overline{u}^{(t-1)})\|^2/(2L) \leq c_t \to 0$. Hence, limit points of $(\theta^{(t)})_{t=1}^T$ are stationary points of J.

Algorithm 1 Relaxed Majorization-Minimization: constant memory version

Require: Initial $\theta^{(0)} = \theta^{(-1)}$ and $\overline{u}^{(0)}$; number of rounds T; $R_{\min} \geq 1$

for $t = 1, \ldots, T$ **do**
  $R \leftarrow R_{\min}$
  repeat
    $J_0 \leftarrow 0$, $J_1 \leftarrow 0$, $g \leftarrow 0$
    for $i = 1, \ldots, N$ **do**
      Set

$$\begin{aligned} &\overline{v} \leftarrow R\text{-step-arg-min}_{\overline{v}} \overline{J}_i(\theta^{(t-1)}, \overline{v}) && \underline{v} \leftarrow R\text{-step-arg-max}_{\underline{v}} \underline{J}_i(\theta^{(t-1)}, \underline{v}) \\ &J_0 \leftarrow J_0 + \overline{J}_i(\theta^{(t-1)}, \overline{v}) && J_1 \leftarrow J_1 + \underline{J}_i(\theta^{(t-1)}, \underline{v}) \\ &g \leftarrow g + \nabla_\theta \overline{J}_i(\theta^{(t-1)}, \overline{v}) \end{aligned} \tag{5}$$

    end for
    $R \leftarrow 2R$
  until $J_0 \leq \eta J_1 + (1 - \eta) J_2$ ▷ Test for the ReGeMM condition
  Set $\theta^{(t)} \leftarrow \theta^{(t-1)} - \frac{1}{L} g$
end for
return $\theta^{(T)}$

Remark 1. $\theta^{(t)}$ is not necessarily induced by a gradient step, but a new iterate $\theta^{(t)}$ has to satisfy a sufficient descent condition,

$$\overline{J}(\theta^{(t)}; \overline{u}^{(t)}) \leq \overline{J}(\theta^{(t-1)}; \overline{u}^{(t)}) - \frac{\kappa \|\nabla \overline{J}(\theta^{(t-1)}; \overline{u}^{(t)})\|^2}{2L}.$$

for a factor $\kappa \in (0, 1)$. In this setting we obtain analogously

$$c_t \geq \frac{\kappa}{2L} \|\nabla \overline{J}(\theta^{(t-2)}; \overline{u}^{(t-1)})\|^2 \geq 0, \tag{4}$$

leading to the same conclusion.

2 ReGeMM using constant memory

As pointed out in the main text, naive implementations of the ReGeMM algorithm (Alg. 1 in the main text) require $O(N)$ memory to store $\overline{\mathbf{u}} = (\overline{u}_1, \ldots, \overline{u}_N)$. In many applications the number of terms N is large, but the latent variables $(\overline{u}_i)_i$ have the same structure for all i (e.g. $\overline{u}_i$ represent pixel-level predictions for training images of the same dimensions). If we use a gradient method to update θ, then the required quantities can be accumulated in-place, in Alg. 1. The constant memory algorithm is not limited to first order methods for θ, but any method that accumulates the information needed to determine $\theta^{(t)}$ from $\theta^{(t-1)}$ in-place is feasible (such as the Newton or the Gauss-Newton method).

The latent variables $\overline{v}$ and $\underline{v}$ are reused for all terms i, and inference for $\overline{v}$ and $\underline{v}$ is conducted using R steps of suitable iterative inference methods, that are

monotonically decreasing and increasing, respectively. Inference is started from constant inital values for $\overline{v}$ and $\underline{v}$. Suitable choices such for inference methods include gradient methods and (block) coordinate descent (if the mappings $\overline{\mathbf{u}} \mapsto \overline{J}(\theta, \overline{\mathbf{u}})$ and $\underline{\mathbf{u}} \mapsto \underline{J}(\theta, \underline{\mathbf{u}})$ are strictly convex and concave, respectively, for all θ). If the ReGeMM condition (Eq. 8 in the main text) is not satisfied, then repeated doubling of R ensures, that the total number of steps spent for inference is at most four times the minimally required number of steps.[1]

The constant memory algorithm is not limited to first order methods to update θ, but any method that accumulates the information needed to determine $\theta^{(t)}$ from $\theta^{(t-1)}$ in-place is feasible (such as the Newton or the Gauss-Newton method).

3 Analysis of stochastic SuDeMM

We use the following assumptions:

1. $\overline{J}_i(\cdot, \overline{u})$ has Lipschitz gradient with constant L for all i and $\overline{u}$. This implies that J has Lipschitz gradient as well.
2. All iterates $\theta^{(t)}$, $t \in \mathbb{N}$, are bounded. Together with the Lipschitz gradient assumption this means, that the sequence of gradients $(g_t)_{t=1}^{\infty}$ is contained in a bounded set.

We partially follow [1]. Let $g_t := \nabla \overline{J}_{i_t}(\theta^{(t-1)}, \overline{u}_{i_t}^{(t)})$ and therefore $\theta^{(t)} = \theta^{(t-1)} - \alpha_t g_t$. Thus, we have

$$\begin{aligned}\overline{J}(\theta^{(t)}, \overline{u}^{(t)}) &= \overline{J}(\theta^{(t-1)} - \alpha_t g_t, \overline{u}^{(t)}) \\ &\leq \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)}) - \alpha_t \nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})^T g_t + \frac{L\alpha_t^2}{2}\|g_t\|^2. \end{aligned} \tag{6}$$

The SuDeMM condition implies that

$$\overline{J}_{i_t}(\theta^{(t-1)}, \overline{u}_{i_t}^{(t)}) \leq J_{i_t}(\theta^{(t-1)}) + \frac{\rho_t}{2}\|g_t\|^2. \tag{7}$$

By taking the expectation on both sides of this relation we consequently obtain

$$\begin{aligned}\mathbb{E}\left[\overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})\right] &\leq J(\theta^{(t-1)}) + \mathbb{E}\left[\frac{\rho_t}{2}\|g_t\|^2\right] \\ &\leq \overline{J}(\theta^{(t-1)}, \overline{u}^{(t-1)}) + \mathbb{E}\left[\frac{\rho_t}{2}\|g_t\|^2\right]. \end{aligned} \tag{8}$$

Combining this with Eq. 6 yields

$$\begin{aligned}\mathbb{E}\left[\overline{J}(\theta^{(t)}, \overline{u}^{(t)})\right] &\leq \overline{J}(\theta^{(t-1)}, \overline{u}^{(t-1)}) + \mathbb{E}\left[\frac{L\alpha_t^2 + \rho_t}{2}\|g_t\|^2 - \alpha_t \nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})^T g_t\right] \\ &= \overline{J}(\theta^{(t-1)}, \overline{u}^{(t-1)}) + \mathbb{E}\left[\frac{L\alpha_t^2 + \rho_t}{2}\|g_t\|^2\right] - \alpha_t \|\nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})\|^2, \end{aligned} \tag{9}$$

[1] In the worst case that the mininum number of steps required to meet the relaxed MM condition is $2^M + 1$ for some $M \in \mathbb{N}$, hence the doubling approach will need $1 + 2 + \cdots + 2^{M+1} \approx 2^{M+2}$ total inference steps.

since $\mathbb{E}\left[g_t\right] = \nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})$. We consider the telescopic sum and obtain

$$\begin{aligned}\mathbb{E}\left[\overline{J}(\theta^{(T)}, \overline{u}^{(T)})\right] - \overline{J}(\theta^{(0)}, \overline{u}^{(0)}) &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\overline{J}(\theta^{(t)}, \overline{u}^{(t)}) - \overline{J}(\theta^{(t-1)}, \overline{u}^{(t-1)})\right)\right] \\ &\leq \sum_{t=1}^{T}\left(-\alpha_t \|\nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})\|^2 + \mathbb{E}\left[\frac{L\alpha_t^2 + \rho_t}{2}\|g_t\|^2\right]\right)\end{aligned}$$

or

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t \|\nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})\|^2\right] &\leq \overline{J}(\theta^{(0)}, \overline{u}^{(0)}) - \mathbb{E}\left[\overline{J}(\theta^{(T)}, \overline{u}^{(T)})\right] \\ &\quad + \sum_{t=1}^{T}\mathbb{E}\left[\frac{L\alpha_t^2 + \rho_t}{2}\|g_t\|^2\right]. \qquad (10)\end{aligned}$$

The r.h.s. is finite by our assumptions. With $\sum_{t=1}^{\infty}\alpha_t = \infty$ this implies that $\liminf_{t\to\infty}\mathbb{E}\left[\|\nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})\|^2\right] = 0$. With the following simple lemma we can show something stronger.

Lemma 1. *Let X_t be a stochastic process adapted to the filtration $(\mathcal{F}_t)_t$ satisfying $\mathbb{E}\left[X_t - \delta_t | \mathcal{F}_{t-1}\right] \leq X_{t-1}$ for all $t \in \mathbb{N}$. Then $Y_t := X_t - \sum_{r=1}^{t}\delta_r$ is a supermartingale.*

Proof. We have

$$\begin{aligned}\mathbb{E}\left[Y_t | \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[X_t - \sum_{r=1}^{t}\delta_r | \mathcal{F}_{t-1}\right] = \mathbb{E}\left[X_t - \delta_t | \mathcal{F}_{t-1}\right] - \sum_{r=1}^{t-1}\delta_r \\ &\leq X_{t-1} - \sum_{r=1}^{t-1}\delta_r = Y_{t-1},\end{aligned}$$

hence Y_t is a supermartingale.

We choose $X_t := \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})$ and

$$\delta_t := \frac{L\alpha_t^2 + \rho_t}{2}\|g_t\|^2 - \alpha_t \|\nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})\|^2. \qquad (11)$$

Using Eq. 9 and the above lemma the stochastic process

$$Y_t := \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)}) - \sum_{r=1}^{t}\frac{L\alpha_r^2 + \rho_r}{2}\|g_r\|^2 + \sum_{r=1}^{t}\alpha_r \|\nabla_\theta \overline{J}(\theta^{(r-1)}, \overline{u}^{(r)})\|^2 \qquad (12)$$

is a supermartingale. By our assumptions on the sequences $(\alpha_t)_t$, $(\rho_t)_t$ and $(g_t)_t$, the sum in the middle is bounded. Further, the first term and the last sum are non-negative. Hence, Y_t is also bounded from below, and via the supermartingale

convergence theorem $Y_t \to Y_\infty < \infty$ a.s. Consequently, the boundedness of $(\overline{J}(\theta^{(t-1)}, \overline{u}^{(t)}))_t$ and $\sum_{r=1}^{\infty} \frac{L\alpha_r^2+\rho_r}{2} \|g_r\|^2$ implies that

$$\sum_{t=1}^{\infty} \alpha_t \|\nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})\|^2 < \infty \qquad \text{a.s.} \tag{13}$$

With $\sum_{t=1}^{\infty} \alpha_t = \infty$ we deduce that $\|\nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)})\|^2 \to 0$ (and therefore $\nabla_\theta \overline{J}(\theta^{(t-1)}, \overline{u}^{(t)}) \to 0$) a.s. Thus, with probability 1 (w.r.t. the sequence of sampled indices $(i_t)_{t=1}^{\infty}$) every accumulation point of $(\theta^{(t)})_{t=1}^{\infty}$ is a stationary point.

4 The list of bundle adjustment instances

The list of the used bundle adjustment datasets is as follows: LADYBUG-73, LADYBUG-138, LADYBUG-318, LADYBUG-598, TRAFALGAR-126, TRAFALGAR-138, TRAFALGAR-201, TRAFALGAR-225, TRAFALGAR-257, DUBROVNIK-150, DUBROVNIK-202, DUBROVNIK-253, DUBROVNIK-308, DUBROVNIK-356, VENICE-89, VENICE-245, VENICE-427, VENICE-744, FINAL-93, FINAL-394.

5 Application to sparse coding

Unsupervised dictionary learning We use the standard sparse coding model as the underlying "network" energy,

$$E(z; x) = \frac{1}{2} \|W_0 z + b_0 - x\|^2 + \kappa \|z\|_1, \tag{14}$$

where $\kappa > 0$ is the weight of the sparsity term. We constrain the columns of W_0 (i.e. the dictionary elements) to have unit norm. For a given dataset $\{x_i\}$ the learning objective is

$$J(W_0, b_0) = \sum_i \min_z \left\{ \frac{1}{2} \|W_0 z + b_0 - x_i\|^2 + \kappa \|z\|_1 \right\} \to \min_{W_0, b_0}. \tag{15}$$

We use coordinate descent to minimize w.r.t. sparse code z, and the standard tangent-plane following by normalization approach to update W_0. J is minimimized w.r.t. W_0 and b_0 using the stochastic variant of SuDeMM. One dual of E is given by

$$E^*(\lambda; x) = -\frac{1}{2} \|\lambda\|^2 + \lambda^T (b_0 - x) \qquad \text{s.t. } \left\|W_0^T \lambda\right\| \le \kappa. \tag{16}$$

Given a primal iterate z, a dual estimate is obtained via $\lambda = W_0 z + b_0 - x$, which is made feasible (if necessary) by uniform scaling of λ,

$$\lambda \leftarrow \frac{\lambda}{\max\{1, \|W_0^T \lambda\|_\infty / \kappa\}}. \tag{17}$$

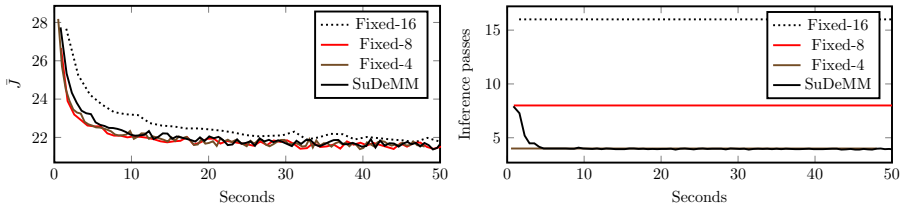

Fig. 1. Unsupervised dictionary learning: objective value w.r.t. wall clock time (left) and the number of inference steps needed to meet the respective criterion (right column) for MNIST using the stochastic gradient method.

In Fig. 1 we depict the evolution of the learning loss and the number of inference steps (i.e. coordinate descent traversals over z) applied, when learning a dictionary (with 128 elements) for the MNIST dataset (with $\kappa = 2$) and the learning rate is constant $\alpha_t = 0.02$. Interestingly, progress in learning leads to more efficient inference, and the number of required inference passes quickly drops for SuDeMM.

Discriminative dictionary learning We stack a linear regression layer on top of the sparse coding layer and arrive at the following network energy,

$$E(z;x) = \frac{1}{2}\left\|W_0 z_1 + b_0 - x\right\|^2 + \kappa\|z_1\|_1 + \frac{\nu}{2}\|z_1\|^2 + \frac{\beta}{2}\left\|W_1 z_1 + b_1 - z_2\right\|^2, \tag{18}$$

where $\beta > 0$ is the feedback weight and $\nu \geq 0$ weights the Tikhonov regularization. Thus, the sparse coding layer becomes an elastic net model. Discriminative learning in such a model is achieved by using the following constrastive loss,

$$\begin{aligned} J(W_0, b_0, W_1, b_1) &= \sum_i \min_z \hat{E}(z; x_i, y_i) - \sum_i \min_z \check{E}(z; x_i) \\ &= \sum_i \min_z \hat{E}(z; x_i, y_i) + \sum_i \max_\lambda -\check{E}^*(\lambda; x_i) \to \min_{W_0, b_0, W_1, b_1}, \end{aligned} \tag{19}$$

where $\{(x_i, y_i)\}$ is a training set. Here the clamped primal and dual energies are given by

$$\begin{aligned} \hat{E}(z;x,y) &= \frac{1}{2}\left\|W_0 z_1 + b_0 - x\right\|^2 + \kappa\|z_1\|_1 + \frac{\nu}{2}\|z_1\|^2 + \frac{\beta}{2}\left\|W_1 z_1 + b_1 - y\right\|^2 \\ \hat{E}^*(\lambda;x,y) &= -\frac{1}{2}\left\|\lambda_1\right\|^2 - \frac{\beta}{2}\left\|\lambda_2\right\|^2 + \lambda_1^T(b_0 - x) + \beta\lambda_2^T(b_1 - y) \\ &\quad - \frac{1}{2\nu}\sum_j \left[|(W_0^T\lambda_1 + \beta W_1^T\lambda_2)_j| - \kappa\right]_+^2, \end{aligned} \tag{20}$$

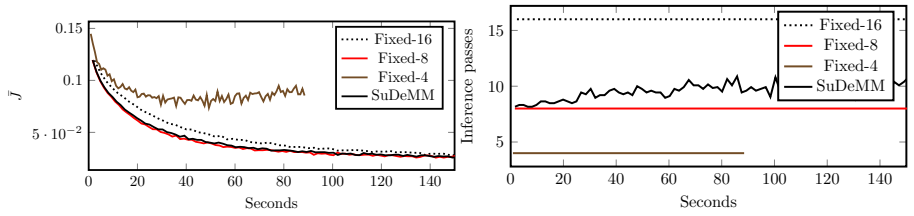

Fig. 2. Discriminative dictionary learning: objective value w.r.t. wall clock time (left) and the number of inference steps needed to meet the respective criterion (right column) for MNIST using the stochastic gradient method.

and the free variants read as

$$\begin{aligned}\check{E}(z;x) &= \frac{1}{2}\left\|W_0 z_1 + b_0 - x\right\|^2 + \kappa\|z_1\|_1 + \frac{\nu}{2}\|z_1\|^2 \\ \check{E}^*(\lambda;x) &= -\frac{1}{2}\left\|\lambda_1\right\|^2 - \frac{\beta}{2}\left\|\lambda_2\right\|^2 + \lambda_1^T(b_0 - x) - \frac{1}{2\nu}\sum_j \left[|(W_0^T\lambda_1)_j| - \kappa\right]_+^2,\end{aligned} \tag{21}$$

Using the elastic net regularizer (with $\kappa = 2$ and $\nu = 1/4$) significantly stablilizes converting between primal and dual variables and therefore improves estimates for the current duality gap. In Fig. 2 we show the evolution of the contrastive learning loss and the number of required inference steps when training on the MNIST dataset. The dictionary W_0 is initialized to the one obtained in the unsupervised setting above. W_1 is initialized with random Gaussian entries, and the initial biases b_0 and b_1 are zero. The learning rate is constant $\alpha_t = 0.02$. We observe a slow (but noisy) increase of necessary inference passes for SuDeMM over time.

References

1. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. Siam Review **60**(2), 223–311 (2018)