Supplementary material for: Convergence Rate Analysis of MAP Coordinate Minimization Algorithms

Ofer Meshi
meshi@cs.huji.ac.il

Tommi Jaakkola
tommi@csail.mit.edu

Amir Globerson
gamir@cs.huji.ac.il

1 Primal Convergence Rate

For clarity, we define

$$\mu \cdot \theta = \sum_i \sum_{x_i} \mu_i(x_i)\theta_i(x_i) + \sum_c \sum_{x_c} \mu_c(x_c)\theta_c(x_c) \tag{1}$$

$$H(\mu) = \sum_i H(\mu_i(\cdot)) + \sum_c H(\mu_c(\cdot)) \tag{2}$$

Theorem 1.1. *Denote by P^*_τ the optimum of the smoothed primal $PMAP_\tau$. Then for any set of dual variables δ, if $\|\nabla F(\delta)\|_\infty \le \epsilon \in R(\tau)$ (for a range of values $R(\tau)$), then $P^*_\tau - P_\tau(\tilde{\mu}) \le C_0\epsilon$, where C_0 is a constant that depends only on the parameters θ, independent of τ, and $\tilde{\mu}$ represents the set of locally consistent marginals from Algorithm 1 in response to $\mu = \mu(\delta)$.*

Proof. $\|\nabla F(\delta)\|_\infty \le \epsilon$ guarantees that $\mu = \mu(\delta)$ are ϵ-consistent in the sense that $|\mu_i(x_i) - \mu_c(x_i)| \le \epsilon$ for all $c, i \in c$ and x_i. Algorithm 1 maps any such ϵ-consistent μ to locally consistent marginals $\tilde{\mu}$ such that

$$|\mu_i(x_i) - \tilde{\mu}_i(x_i)| \le 3\epsilon N_{\max}, \quad |\mu_c(x_c) - \tilde{\mu}_c(x_c)| \le 2\epsilon N^2_{\max}, \tag{3}$$

for all i, x_i, c, and x_c, where $N_{\max} = \max\{\max_i N_i, \max_c N_c\}$. In other words, $\|\mu - \tilde{\mu}\|_\infty \le K\epsilon$. This can be easily derived from the update in Algorithm 1 and the fact that $|\mu_i(x_i) - \mu_c(x_i)| \le \epsilon$.

Next, it can be shown that $F(\delta) = P_\tau(\mu(\delta))$. And it follows that $P^*_\tau \le F(\delta) \le P_\tau(\mu)$, where the first inequality follows from weak duality.

Thus we have:

$$P^*_\tau \le P_\tau(\mu) = \mu \cdot \theta + \frac{1}{\tau}H(\mu) = (\tilde{\mu} + \mu - \tilde{\mu}) \cdot \theta + \frac{1}{\tau}H(\tilde{\mu}) + \frac{1}{\tau}(H(\mu) - H(\tilde{\mu})) \tag{4}$$

$$\le P_\tau(\tilde{\mu}) + \|\mu - \tilde{\mu}\|_\infty \|\theta\|_1 + \frac{1}{\tau}(H(\mu) - H(\tilde{\mu})) \tag{5}$$

$$\le P_\tau(\tilde{\mu}) + K\epsilon\|\theta\|_1 + \frac{1}{\tau}(H(\mu) - H(\tilde{\mu})) \tag{6}$$

Where we have used Holder's inequality for the first inequality and Eq. (3) for the second inequality.

It remains to bound $\frac{1}{\tau}(H(\mu) - H(\tilde{\mu}))$ by a linear function of ϵ. We note that it is impossible to achieve such a bound in general (e.g., see [1]). However, since the entropy is bounded the difference is also bounded. Now, if we also restrict ϵ to be large enough $\epsilon \ge \frac{1}{\tau}$, then we obtain the bound:

$$\frac{1}{\tau}(H(\mu) - H(\tilde{\mu})) \le \frac{1}{\tau}H_{\max} \le \epsilon H_{\max} \tag{7}$$

We thus obtain that Eq. (6) is of the form $P_\tau(\tilde{\mu}) + O(\epsilon)$ and the result follows.

For the high-accuracy regime (small ϵ) we provide a similar bound for the case $\epsilon \leq O(e^{-\tau})$. Let $v = \mu - \tilde{\mu}$, so we have:

$$\begin{aligned} H(\mu) - H(\tilde{\mu}) &= H(\tilde{\mu} + v) - H(\tilde{\mu}) \\ &\leq H(\tilde{\mu}) + \nabla H(\tilde{\mu})^\top v - H(\tilde{\mu}) \\ &= -\sum_i \sum_{x_i} v_i(x_i) \log \tilde{\mu}_i(x_i) - \sum_c \sum_{x_c} v_c(x_c) \log \tilde{\mu}_c(x_c) \end{aligned}$$

where the inequality follows from the concavity of entropy, and the second equality is true because $\sum_{x_i} v_i(x_i) = 0$ and similarly for $v_c(x_c)$. Now, from the definition of $\mu_i(x_i;\delta)$ we obtain the following bound:

$$\mu_i(x_i;\delta) = \frac{1}{Z_i} e^{\tau(\theta_i(x_i) + \sum_{c:i\in c} \delta_{ci}(x_i))} \geq \frac{1}{|X_i|} e^{-2\tau(\|\theta_i\|_\infty + \|\delta_i\|_1)}$$

We will show below (Lemma 1.2) that $\|\delta_i\|_1$ remains bounded by a constant A independent of τ. Thus we can write:

$$\mu_i(x_i;\delta) \geq \frac{1}{|X_{\max}|} e^{-2\tau(\|\theta_i\|_\infty + A)}$$

where $|X_{\max}| = \max\{\max_i |X_i|, \max_c |X_c|\}$. We define $\gamma_0 = \frac{1}{(2|X_{\max}|)^\tau} e^{-2\tau(\|\theta_i\|_\infty + A)}$, and thus for any $\tau \geq 1$ we have that $\mu_i(x_i;\delta)$ is bounded away from zero by $2^\tau \gamma_0$. Since we assume that $\epsilon \leq \gamma_0$, we can bound $\tilde{\mu}$ from below by γ_0. As a result, since $\|v_i\|_\infty \leq K\epsilon$,

$$-\frac{1}{\tau} \sum_i \sum_{x_i} v_i(x_i) \log \tilde{\mu}_i(x_i) \quad \leq \quad -\frac{1}{\tau}(\log \gamma_0)|X_i|K\epsilon = (2(\|\theta_i\|_\infty + A) + \log(2|X_{\max}|))|X_i|K\epsilon$$

and similarly for the other entropy terms.

Again, we obtain that Eq. (6) is of the form $P_\tau(\tilde{\mu}) + O(\epsilon)$ and the result holds.

In conclusion, we have shown that if $\|\nabla F(\delta)\|_\infty \leq \epsilon$, then for large values $\epsilon \geq \frac{1}{\tau}$ and small values $\epsilon \leq \frac{1}{(2|X_{\max}|)^\tau} e^{-2\tau(\|\theta_i\|_\infty + A)}$ we have that: $P_\tau^* - P_\tau(\tilde{\mu}) \leq O(\epsilon)$. Our analysis does not cover values in the middle range, but we next argue that the covered range is useful. □

The allowed range of ϵ (namely $\epsilon \in R(\tau)$) seems like a restriction. However, as we argue next taking $\epsilon \geq \frac{1}{\tau}$ (i.e., $\epsilon \in R(\tau)$) is all we need in order to obtain a desired accuracy in the non-smoothed primal.

Suppose one wants to solve the original problem $PMAP$ to within accuracy ϵ'. There are two sources of inaccuracy, namely the smoothing and suboptimality. To ensure the desired accuracy, we require that $P_\tau^* - P^* \leq \alpha\epsilon'$ and likewise $P_\tau(\tilde{\mu}) - P_\tau^* \leq (1-\alpha)\epsilon'$. In other words, we allow $\alpha\epsilon'$ suboptimality due to smoothing and $(1-\alpha)\epsilon'$ due to suboptimality.

For the first condition, it is enough to set the smoothing constant as: $\tau \geq \frac{H_{\max}}{\alpha\epsilon'}$. The second condition will be satisfied as long as we use an ϵ such that: $\epsilon \leq \frac{(1-\alpha)\epsilon'}{(K\|\theta\|_1 + H_{\max})}$ (see Eq. (6) and Eq. (7)). If we choose $\alpha = \frac{H_{\max}}{K\|\theta\|_1 + 2H_{\max}}$ we obtain that this ϵ satisfies $\epsilon \geq \frac{1}{\tau}$ and therefore $\epsilon \in R(\tau)$.

Lemma 1.2. *Assume δ is a set of dual variables satisfying $F(\delta) \leq F(0)$ where $F(0)$ is the dual value corresponding to $\delta = 0$. We can require $\sum_{c:i\in c} \delta_{ci}(x_i) = 0$ since $F(\delta)$ is invariant to constant shifts. Then it holds that:*

$$\sum_{c,i,x_i} |\delta_{ci}(x_i)| = \|\delta\|_1 \leq A \tag{8}$$

where

$$A = 2\max_i |X_i| \left(F(0) + \sum_i \max_{x_i} |\theta_i(x_i)| + \sum_c \max_{x_c} |\theta_c(x_c)| \right) \tag{9}$$

Proof. To show this, we bound

$$\begin{aligned} &\max_{\delta} \sum_{c,i,x_i} r_{ci}(x_i)\delta_{ci}(x_i) \\ \text{s.t. } & F(\delta) \leq F(0) \\ & \sum_{c:i\in c} \delta_{ci}(x_i) = 0 \end{aligned} \tag{10}$$

For any $r_{ci}(x_i) \in [-1, 1]$. The dual problem turns out to be:

$$\begin{array}{ll} \min\limits_{\mu,\gamma,\alpha} & \alpha(F(0) - \sum\limits_{c,x_c} \mu_c(x_c)\theta_c(x_c) - \sum\limits_{i,x_i} \mu_i(x_i)\theta_i(x_i) - \sum\limits_{i} H(\mu_i(x_i)) - \sum\limits_{c} H(\mu_c(x_c)) \\ \text{s.t.} & \mu_i(x_i) - \mu_c(x_i) = \frac{r_{ci}(x_i)-\gamma_{ci}}{\alpha} \\ & \mu_i(x_i) \geq 0, \mu_c(x_c) \geq 0 \\ & \sum_{x_i} \mu_i(x_i) = 1, \sum_{x_c} \mu_c(x_c) = 1 \\ & \alpha \geq 0 \end{array} \tag{11}$$

We will next upper bound this minimum with a constant independent of r and thus obtain an upper bound that holds for all r. To do this, we will present a feasible assignment to the variables α, μ, γ above and use the value they attain. First, we set $\alpha = \hat{\alpha} = 2\max_i |X_i|$. Next, we note that for this $\hat{\alpha}$, the objective of Eq. (11) is upper bounded by A (as defined in Eq. (9)). Thus we only need to show that $\hat{\alpha} = 2\max_i |X_i|$ is indeed a feasible value, and this will be done by showing feasible values for the other variables denoted by $\hat{\mu}, \hat{\gamma}$. First, we set:

$$\hat{\mu}_i(x_i) \quad = \quad \frac{1}{|X_i|}$$

and:

$$\hat{\gamma}_{ci} = \frac{1}{|X_i|} \sum_{x_i} r_{ci}(x_i) \tag{12}$$

Next, we define $\nu_{ci}(x_i)$ (for all c, i, x_i) as follows:

$$\nu_{ci}(x_i) = \hat{\mu}_i(x_i) - \frac{r_{ci}(x_i) - \hat{\gamma}_{ci}}{\hat{\alpha}} \tag{13}$$

It can easily be shown that $\nu_{ci}(x_i)$ is a valid distribution over x_i (i.e., non negative and sums to one). Thus we can define:

$$\hat{\mu}_c(x_c) = \prod_{i\in c} \nu_{ci}(x_i) \tag{14}$$

Since $\hat{\mu}_c(x_c)$ is a product of distributions over the variables in c, it is also a valid distribution. Thus it follows that all constraints in Eq. (11) are satisfied by $\hat{\alpha}, \hat{\gamma}, \hat{\mu}$, and the desired bound holds.

□

2 Star improvement bound

We prove the following proposition:

Proposition 2.1. *The star update for variable x_i satisfies:*

$$F(\delta^t) - F(\delta^{t+1}) \geq \frac{1}{4\tau N_i} \|\nabla_{S_i} F(\delta^t)\|_2^2$$

Proof. First, we know that the improvement associated with the star update for variable x_i is:

$$F(\delta^t) - F(\delta^{t+1}) = -\frac{1}{\tau} \log \left(\sum_{x_i} \left(\mu_i^t(x_i) \cdot \prod_{c:i\in c} \mu_c^t(x_i) \right)^{\frac{1}{N_i+1}} \right)^{N_i+1}$$

Therefore, for any probability distributions $p, q^{(1)}, ..., q^{(m)}$ we want to prove that:

$$\left(\sum_i \left(p_i \cdot \prod_k q_i^{(k)} \right)^{\frac{1}{m+1}} \right)^{m+1} \quad \leq \quad \exp\left(-\frac{1}{4m} \sum_k \sum_i \left(p_i - q_i^{(k)} \right)^2 \right)$$

Lemma 2.2. *For any probability distributions $p, q^{(1)}, ..., q^{(m)}$ the following holds:*

$$\left(\sum_i \left(p_i \cdot \prod_k q_i^{(k)}\right)^{\frac{1}{m+1}}\right)^{m+1} \leq 1 - \frac{1}{4m}\sum_k \left(\sum_i |p_i - q_i^{(k)}|\right)^2$$

Proof.

$$\begin{aligned}
\sum_k \left(\sum_i |p_i - q_i^{(k)}|\right)^2 &\leq \sum_k \left(\sum_i (\sqrt{p_i} - \sqrt{q_i^{(k)}})^2 \cdot \sum_i (\sqrt{p_i} + \sqrt{q_i^{(k)}})^2\right) \\
&= \sum_k \left(4 - 4\left(\sum_i \sqrt{p_i q_i^{(k)}}\right)^2\right) \\
&= 4m - 4\sum_k \left(\sum_i \sqrt{p_i q_i^{(k)}}\right)^2 \\
&\leq 4m - 4\sum_k \left(\sum_i \left(p_i \cdot \prod_{k'} q_i^{(k')}\right)^{\frac{1}{m+1}}\right)^{m+1} \\
&= 4m - 4m\left(\sum_i \left(p_i \cdot \prod_{k'} q_i^{(k')}\right)^{\frac{1}{m+1}}\right)^{m+1} \\
\Rightarrow \left(\sum_i \left(p_i \cdot \prod_{k'} q_i^{(k')}\right)^{\frac{1}{m+1}}\right)^{m+1} &\leq 1 - \frac{1}{4m}\sum_k \left(\sum_i |p_i - q_i^{(k)}|\right)^2
\end{aligned}$$

For the first transition see [3] (also in [2] p. 57). The second inequality follows from Theorem 1 in [4]. □

Using Lemma 2.2 the desired result follows since:

$$\begin{aligned}
\left(\sum_i \left(p_i \cdot \prod_k q_i^{(k)}\right)^{\frac{1}{m+1}}\right)^{m+1} &\leq 1 - \frac{1}{4m}\sum_k \left(\sum_i |p_i - q_i^{(k)}|\right)^2 \\
&\leq 1 - \frac{1}{4m}\sum_k \sum_i \left(p_i - q_i^{(k)}\right)^2 \\
&\leq \exp\left(-\frac{1}{4m}\sum_k \sum_i \left(p_i - q_i^{(k)}\right)^2\right)
\end{aligned}$$

□

3 Gradient-based algorithms

In this section we describe the gradient descent and FISTA algorithms used in the experiments.

Algorithm 1: Gradient descent

for $t = 1, \ldots$ **do**
$\quad \delta^{t+1} = \delta^t - \frac{1}{L}\nabla F(\delta^t)$
end for

Algorithm 2: FISTA

$\bar{\delta}^1 = \delta^0, \quad \alpha^1 = 1$
for $t = 1, \ldots$ **do**
$\quad \delta^t = \bar{\delta}^t - \frac{1}{L}\nabla F(\bar{\delta}^t)$
$\quad \alpha^{t+1} = \frac{1+\sqrt{1+4(\alpha^t)^2}}{2}$
$\quad \bar{\delta}^{t+1} = \delta^t + \left(\frac{\alpha^t - 1}{\alpha^{t+1}}\right)\left(\delta^t - \delta^{t-1}\right)$
end for

References

[1] D. Berend and A. Kontorovich. A reverse pinsker inequality. *CoRR*, abs/1206.6544, 2012.

[2] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *Communication Technology, IEEE Transactions on*, 15(1):52 –60, february 1967.

[3] C. Kraft. Some conditions for consistency and uniform consistency of statistical procedures. In *Univ. of California Publ. in Statistics, vol. 1*, pages 125–142. Univ. of California, Berkeley, 1955.

[4] K. Matusita. On the notion of affinity of several distributions and some of its applications. *Annals of the Institute of Statistical Mathematics*, 19:181–192, 1967. 10.1007/BF02911675.